2018-09-20

What we did

- James**:**
- **Jim**: tensorly/sparse work
- Martin: many-parquet-files, JIRAs for pyarrow parquet
- Matt: incremental model selection, docs, handling external issue/prs
- Scott:
- Tom: pandas CI, arrays

Agenda items

- Documentation discussion (update)
    - Need to update a bunch of links
    - Probably want to wait until Andy/NumFOCUS enables https
    -
- Restructure Meeting
    - Matt's initial plan:
        - Have an anaconda employees only daily standup around a kanban board, aim for 10-15 minutes. Replace daily "Today" messages in gitter chat. Maybe downgrade to every MWF if things seem like it's too much.
        - Use github projects for kanban board
        - Move community meeting to monthly and schedule 60-90 minutes. Focus more around larger issues than on status updates.
- Blog.dask.org -> dask.github.io/blog
    - People don't seem to have much preference around blogging technology
    - Matt will copy his current setup
- Avro and other data types

Pending Work

- Incremental hyper-parameter optimization
    - Hyperband (with random search)
- Parameter Servers
- Task fusion
    - Fuse Dataframes - Erik has reviewed, planning to finish up today/tomorrow
    - Atop fusion - Matt to take this on
- Experiment with jupyter-enterprise-gateway and Hadoop clusters
    - Test cluster set up

    - Ongoing conversation
- Documentation (work ongoing)
- Deployment (no major activity planned for this week)
    - Plenty of dask-jobqueue PRs to review
    - Organizing meeting between groups
- Evangelism
    - Official Dask blog
- Scheduling performance
    - Non-tornado Comms (stalled)
- 1.0 (no work planned this week)

What we plan to do and availability

- Derek: Pangeo, other consulting work
- James:
- Jim: Finish task fusion, continue on sparse/tensorly work
- Martin: unsure, may be helping out with consulting work; avro readers
- Matt: atop fusion, fix distributed scheduling intermittent errors, dask blog
- Scott:
- Tom: Off Friday - Tuesday.

2018-09-13

What we did

- James**:**
- Jim: Off Friday/Monday for wedding, Merged Skein WebUI (https://github.com/jcrist/skein/pull/68), WIP PR for task fusion, responded to PRs/issues
- Martin: multiple paths to fastparquet; transform for Gaia DR2; fastparquet V pyarrow (issue); intake
- Matt: supported incremental model selection and docs. Doing anaconda inc booth work at conference.
- Mike: landsat spectral clustering example
- Scott:
- **Tom**: Model selection, docs, pandas

Agenda items

- Migrate default from fastparquet to pyarrow (update)
    - Made FP issue, invited wesm to make jiras - may be that some items are already covered

- Documentation discussion (update)
    - Dask-examples needs a static page, looking at nbsite (from pyviz folks) and nbsphinx (currently used by dask-ml). Matt is inclined towards nbsphinx currently because it is more stable, but hasn’t actually gotten things to work nicely yet (help welcome)
    - New repo: https://github.com/dask/dask.github.io for landing page at dask.org
        - Just HTML & CSS for now. Will move to static site generator
    - Will move http://dask.pydata.org/en/latest/docs.html to docs.dask.org
    - https?

Pending Work

- Incremental hyper-parameter optimization
    - Hyperband (with random search)
    - Stop on plateau (with random search)
- Parameter Servers
- Task fusion
    - Fuse Dataframes - WIP PR up (https://github.com/dask/dask/pull/3979)
    - Atop fusion - Jim to provide estimate by next week
- Experiment with jupyter-enterprise-gateway and Hadoop clusters
    - Ongoing conversation
- Documentation (work ongoing)
    - Move subprojects to dask.org
    - Include dask-example subfolders in sphinx docs
- Deployment (no major activity planned for this week)
    - Plenty of dask-jobqueue PRs to review
    - Organizing meeting between groups
- Evangelism
    - Official Dask blog (blocked by documentation)
- Scheduling performance
    - Tornado PR: https://github.com/tornadoweb/tornado/pull/2469 (done, needs review)
    - Non-tornado Comms (stalled)
- 1.0 (no work planned this week)

What we plan to do and availability

- James:
- Jim: Finish task fusion PR, start back on catop fusion
- Martin:
- Matt:
- Mike: intake-xarray (opendap support), datashader regrid bug
- Scott:

- Tom: Dask-ML pipelines and performance, blogging ColumnTransformer

2018-09-06

What we did

- James**:**
- Jim: (long weekend), Skein WebUI, meetings, reviewed tensorly work, responded to some user issues.
- Martin: (longer weekend) still intake; zarr/gcsfs discussions; fastparquet fixes
- Matt: Reviewed and pushed along many small bugfixes. Published bugfix release. Some blogpost writing.
- **Mike**: Intake, intake-xarray, Intake caching blog post, landsat multiband spectral clustering with dask-ml
- Scott: Running experiments on model selection
- Tom: dask doc update, incremental

Agenda items

- Migrate default from fastparquet to pyarrow (update)
  - Still need to submit a JIRA
- Documentation discussion (update)
  - NumFOCUS has dask.org
  - One option is to use subdomains like ml.dask.org, yarn.dask.org . Do we know how to do dask.org/ml/…?
  - Need consistent theme across documentation
    - Global navbar at the top that navigates between projects
    - Sidebar to help navigate within a particular project
    - github.com/dask/dask-sphinx-theme
- Project management (update?)
  - We're deciding between something lightweight, like the entry below, or something like a kanban board. We'll try things out this week and report back next.
  - Github projects < ZenHub < Jira in terms of weight

Pending Work

- Incremental hyper-parameter optimization
  - Hyperband (with random search)

    - Stop on plateau (with random search)
- Parameter Servers
- Task fusion
    - Fuse Dataframes - Jim thinks that this should be done by next week
    - Atop fusion - Jim to provide estimate by next week
- Experiment with jupyter-enterprise-gateway and Hadoop clusters
    - Ongoing conversation
- Documentation (work ongoing)
    - Move subprojects to dask.org
    - Include dask-example subfolders in sphinx docs
- Deployment (no major activity planned for this week)
    - Plenty of dask-jobqueue PRs to review
    - Organizing meeting between groups
- Evangelism
    - Official Dask blog (blocked by documentation)
- Scheduling performance
    - Tornado PR: https://github.com/tornadoweb/tornado/pull/2469 (done, needs review)
    - Non-tornado Comms (stalled)
- 1.0 (no work planned this week)

What we plan to do and availability

- James:
- Jim: Out Friday (and maybe Monday) for a wedding, continue task fusion, coordinate with other deployment projects
- Martin: available for small things
- Matt: Documentation, Helm charts, Conference. At conference Tu-Th so will be unresponsive.
- Mike: multiband spectral clustering, intake
- Scott: run experiments, write blog post
- Tom: Incremental Search base, doc PRs to xgboost & scikit-learn, blog column transformer

2018-08-30

What we did

- James**:**
- Jim: Sick, started on task fusion, reviews
- Martin: still intake

- **Matt**: Vacation time, Release prep, blogpost draft on institutional involvement in OSS (please do not publish), helping with dashboard work, and bugfixes
- Mike: intake, landsat multiband spectral clustering with dask-ml
- Scott: Hyperparameter optimization: running simulations + generating comparisons
- Tom: Dask-XGBoost, dask & tpot, dask-ml release

Agenda items

- Dask.org moving slowly from Anaconda to NumFOCUS
- Any ideas on how to get more value out of dask-examples?
    - Integrate dask-examples into sphinx example documentation
    - PyViz generates nicely styled web pages from rendered notebooks
    - Do we want to strip outputs? Dask-ml runs the notebooks during the build process.
- Migrate default from fastparquet to pyarrow
    - This is less clear.
    - A Dask issue linking to all the Arrow JIRAs required to make the move may be helpful
- Migrate default from hdfs3 to pyarrow
    - Martin thinks that HDFS3 should be retired. Too many holes, and no longer provides anything extra that people use. Exceptions: you can access the cluster externally from the cluster without config files.
    - TODO:
        - Change internal dask default
        - Add note on repository
- Project management
    - We're deciding between something lightweight, like the entry below, or something like a kanban board. We'll try things out this week and report back next.

Pending Work

- Release
- Incremental hyper-parameter optimization
    - Hyperband (with random search)
    - Stop on plateau (with random search)
- Parameter Servers
- Task fusion
    - Fuse Dataframes?
    - Atop fusion-
- Experiment with jupyter-enterprise-gateway and Hadoop clusters
- Documentation (pending on domain availability)
    - Move subprojects to dask.org
    - Include dask-example subfolders in sphinx docs

- Deployment
    - Organizing meeting between groups
    - Plenty of dask-jobqueue PRs to review
- Evangelism
    - Official Dask blog?
    - Downstream documentation PRs
    - User survey
- Scheduling performance
    - Tornado PR: https://github.com/tornadoweb/tornado/pull/2469
    - Non-tornado Comms
- 1.0

What we plan to do and availability

- James:
- Jim: continue task fusion, read/review Scott's tensorly work, coordinate with other deployment projects
- Martin: some time off
- Matt: Release, deployment, some blogpost writing
- Mike: multiband spectral clustering, intake blog
- Scott: internship ending friday :(. My plans include seeing Hyperband through to completion (merge PR, blog post, maybe an async variant).
- Tom: Off Monday. Dask-XGBoost, blogging tpot example, submitting doc PRs to xgboost, scikit-learn, pandas sparse

2018-08-23

What we did

- James**:**
- Jim: https://jcrist.github.io/venv-pack/, responded to PRs/issues in tiny yarn ecosystem projects, meetings, started back on task fusion for dask
- **Martin**: only intake
- Matt: Cleaned up and merged PRs from previous weeks. Ready-ish to release now?.
- Mike: some intake; client work
- Scott: hyperparam optimization: compared successive halving and "stop on plateau" with realistic example

- Tom: pandas SparseArray, dask.dataframe compatibility, TPOT

Agenda items

- Raise any release blockers now (dask, distributed)

What we plan to do and availability

- James:
- Jim: Vacation tomorrow -> Monday, task fusion, maybe another blogpost
- Martin: intake
- Matt: Vacation until Monday. Release. Writing some community blogposts.
- Mike: client work
- Scott: implement hyperband with _incremental.fit
- Tom: Possibly dask-xgboost.

2018-08-16

What we did

- James**:**
- **Jim:** Blogposts, fixed a few user issues, PR reviews, conference proposals
- Martin: intake gui
- Matt: Pangeo meeting, banged on performance optimization PRs
- Mike: intake caching.
- Scott: More integration of example into hyperparameter comparison, some work on scaling out parameter server (larger model, better metrics)
- Tom: distributed TPOT: https://github.com/EpistasisLab/tpot/pull/730, pandas array, proposal for dask tutorial at ODSC West accepted

Agenda items

- From last week: Dask ML roadmap? https://github.com/dask/dask-ml/pull/322, https://www.quansight.com/projects
  - Will keep the “values” section and link to a GH issues tag
- Pangeo meeting highlights
  - XArray + rasterio improvements, both docs and some functionality, focused around geotiffs: https://github.com/pydata/xarray/pull/2255

  - Got a few scientists using XArray + Dask on various clusters
  - Dask enabled Binder
  - Dask JupyterLab extension
- Dask.org for docs and moving away from RTD
  - Will check on dask.org transfer, experimenting with redirects

What we plan to do and availability

- James:
- Jim: Continue yarn evangelism. I have time for Dask work, but nothing specific in mind? (atop fusion?)
- Martin: still in intake land
- Matt: Back to working on performance issues and iterating on ML things with Scott
- Mike: revisiting Tom's sklearn xarray dask-ml landsat example
- Scott: Finish hyperparameter comparison and start implementing
- Tom: Finish TPOT, more pandas, Maybe half time tomorrow.

2018-08-09

What we did

- **James:** Summer school in Italy
- Jim**:** Skein docs, new release, a few dask bugs, blogposts
- Martin: zarr conversation (PR) stalled since A Miles away; some time off; poking people about Intake and xarray streaming
- Matt: Investigated performance of scheduler and workers. Resulted in many small optimizations and overhead reductions. No great value to ML workloads yet though.
- Mike: Intake Caching PR
- Scott: Built out hyperparameter example and parameter server (including drafts and timings)
- Tom: distributed TPOT benchmarking

Agenda items

- From last week: Meetings issue now on github https://github.com/dask/dask/issues/3854
- From last week: Docs hosting / restructuring, Peter Wang seems to be ok with giving dask.org to NumFOCUS and having us manage it separately from PyData.
  - dask.pydata.org/yarn ?

    - dask.pydata.org/en/latest/yarn ?
- Do we want to include intake stuff in this meeting?
- Dask ML roadmap? https://github.com/dask/dask-ml/pull/322, https://www.quansight.com/projects

What we plan to do and availability

- James: Catch up on lingering issues / PRs
- Jim: Finish up blogposts, examples, dask issues (nothing specific in particular, if anyone has needs)
- Martin: some time off, mostly Intake
- Matt: Out Friday on vacation. At Pangeo meeting next week. Will probably also clean up some performance-optimization work.
- Mike: Intake Caching (wrap up PR and blog), intake-xarray, examples for client, dask.bytes.HTTPFileSystem
- Scott: Out tomorrow for moving :(
- Tom:

2018-08-02

What we did

- James**:**
- Jim: Finished up skein key-value store, updated dask-yarn to new skein api, released skein and dask-yarn, documentation, ramping back up on dask stuff
- Martin: intake (release blog)
- Matt:
    - Clean up intermittent failures on dask/distributed prior to release.
    - Look at the performance of the event loop thread (https://github.com/dask/distributed/pull/2144)
    - Small issues maintenance
- Mike:
- Scott: Performance benchmarks on successive halving, implementation and performance of parameter servers
- **Tom**: dask-xgboost with multiple GPUs, dask-ml docs, column-transformer

Agenda items

- Rethinking this meeting and general project management process
    - It's nice to get a sense of what's going on
    - It seems like we don't get non-anaconda folks into the meetings

    - Things we could do
        - Occasional longer meetings, explicitly invite a larger crowd
        - More focused meetings, like the dask-ml meeting
        - Should we move this conversation to github
    - Hard to enforce project management on a community project
    - Project management might be necessary at some scale
- Docs hosting / restructuring
    - It might be nice to push subprojects to dask.pydata.org/foo like dask.pydata.org/ml dask.pydata.org/yarn, etc..
    - Anaconda Inc appears to own dask.org, could do ml.dask.org, yarn.dask.org

What we plan to do and availability

- James:
- Jim: Skein documentation, yarn blogpost, ramp back up on dask stuff
- Martin: zarr consolidation? Intake feedback. Release s3fs, fastparquet
- Matt: release dask.distributed, investigate performance around some ML things like actors and incremental model selection
- Mike: off platform team finally
- Scott: More parameter server development, polish successive halving.
- Tom: mostly pandas

2018-07-26

What we did

- James**:**
- Jim: Mostly vacation
- Martin: Intake
- **Matt**: Conference, stateful processing with Actors, small issues
- Mike:
- Scott: Started to experiment with Hyperband and incremental._fit
- Tom: Tabular data workloads (LabelEncoder uses categoricals, OneHotEncoder, ColumnTransformer), tpot experiment

Agenda items

- KubeFlow: Does anyone have experience?

What we plan to do and availability

- James:
- Jim: Finish up Skein concurrency primitives (https://github.com/jcrist/skein/pull/40), release skein & dask-yarn with updated api, docs
- Martin: intake https://github.com/martindurant/intake-release-blog
- Matt: personal time, blogposts,
- Mike:
- Scott: draft successive halving performance (w/ random sampling), scale up one example
- Tom: Case studies.

2018-07-19

What we did

- James**:** SciPy 2018
- Jim:
- **Martin**: scipy events
- Matt: Prototyped a bunch of things:
  - Automatic machine learning with TPOT: https://github.com/EpistasisLab/tpot/pull/730
  - Can we meaningfully help scikit-optimize other than straight joblib: https://github.com/dask/dask-ml/issues/300
  - Possible solution for incremental model selection: https://github.com/dask/dask-ml/pull/288
  - Dataframe examples + videos: https://github.com/dask/dask-examples/pull/24 (Albert looking at this today I think)
  - Also working on actors a bit: https://github.com/dask/distributed/issues/2109
- Mike: AE 5.2 testing, Intake caching
- Scott: Examples (Criteo, Anaconda conda install)
- Tom: Played with tensorflow datasets API, played with delayed fits for dask-ml, catching up on issues.

Agenda items

- Scipy roundup

    - Tutorial went well. Students seemed engaged. Cluster was dead, but easily restarted. Good engagement from students.
    - http://matthewrocklin.com/blog/work/2018/07/17/dask-dev
    - Talks, BoF, sprints
    - Surprising number of astronomers at sprints
- Demos for companies
    - Quick slide deck
    - https://mybinder.org/v2/gh/dask/dask-examples/master?filepath=dataframe.ipynb
- Conda install dataset

What we plan to do and availability

- James: Finish up dask-ml #280 and debug HTCondorCluster for dask-jobqueue
- Jim:
- Martin: intake blog finally; pangeo systems/updating?
- Matt: Next week at conference and with family in California
- Mike: AE 5.2 testing, Intake caching, familiarizing myself with dask-jobqueue
- Scott: Scale Criteo example to large data, cluster conda install dataset example.
- Tom: Follow up on Dask-ML emails and PRs. Maybe dask-xgboost with distributed.

2018-07-12

(I think we’re all busy at SciPy)

What we did

- James**:**
- **Jim**:
- Martin:
- Matt:
- Mike:
- Scott:
- Tom:

Agenda items

-

What we plan to do and availability

- James:
- Jim:
- Martin:
- Matt:
- Mike:
- Scott:
- Tom:

2018-06-28

What we did

- Jim:
- Martin: playing with zarr optimization, pangeo help and talk about examples
- Matt: Task ordering fixes for map-overlap and dask-ml. Cleaned up some dask-jobqueue work, Joe Hamman released. Fooling around a bit with large images with XArray. Also fooling around with HPC systems.
- Tom: SpectralClustering performance
- Mike:
- **Scott**: Re-implemented HyperbandCV, now ready for merge.
- James: Added LabelEncoder to dask-ml

Agenda items

- Dask-SearchCV -> Dask-ML? https://github.com/dask/dask-searchcv/issues/73
  - Probably waiting on jcrist for feedback
  - What about other packages? (xgboost, tensorflow, glm)
- Thoughts on auto-formatters: https://github.com/dask/dask-ml/pull/237
  - Trying using black
  - Removes opinion and conversation on styling
- Scipy tutorial
  - Some changes to infrastructure in separate branch for scikit-learn/image sprint Tom should merge back in
  - Graphviz still a pain
  - Some issues still exist in dask-tutorial repository
- SciPy sprints
  - Label existing issues
  - Add examples issues

    - Finding Documentation
- Dask-jobqueue and Condor
    - James and Scott are interested in this. They happen to be next to the condor people.

What we plan to do and availability

- Jim: Dask-yarn, probably off a bit for medical stuff
- Martin: more pangeo, tutorial infrastructure, issue labeling
- Matt: Focusing on blogposts, maybe large images, maybe HPC. Handling issues.
- Mike:
- Tom: Dask-ML Roadmap, skimage / ml / dask blogpost, Pandas dev sprint print. Traveling starting next Thursday.
- Scott: Some work on examples, getting started with tensor decomp. Time off on the 4th of July.
- James: Look into adding more scoring metrics for dask-ml. Probably not doing much though.

2018-06-21

What we did

- Jim: Off for medical stuff early this week. Dask-Yarn proof of concept is up, will need some more work. Released Skein 0.0.3 and pushed to PyPI and conda-forge.
- Martin:setting up pangeos, arrow-parquet, fastparquet bug, requests bug. Working on splitting off pangeo examples to a separate repository.
- Matt: Released everything. Client meetings. Issue triage and PR review. Some dask array fixes for map-overlap (correctness), and slicing with array-of-ints (performance for xarray).
- **Tom:** Off Friday & Tuesday. Webinar prep (today). Dask / distributed debugging. Dask-ML release.
- Mike:
- Scott: Worked on Hyperband example, received useful feedback from Matthew. Some tensor decomposition work.
- James: Documentation fixups. Dask.array.average (Dask-ML would like to use it)

Agenda items

- Bugfix release for dask/dask?
- From last week:
    - Engage user and developer communities

        - Advanced User Survey: https://docs.google.com/forms/d/e/1FAIpQLSexQY6Y_RZcpkiUOOfyyXns5uoi56E7cL2NDf7HLYoLtjMZWQ/viewform
        - Can it become more obvious that these are possible paragraphs
        - Can we combine this with a general user survey
        - Solicit links for existing blogposts
        - Might want to follow up with additional e-mail asking for official logo
    - SciPy sprints
        - https://github.com/dask/dask/issues/3614
        - Registered!
        - Prep? No updates or effort so far.
- FYI: Pandas dev sprint at Anaconda offices July 5th-8th

What we plan to do and availability

- Jim: Dask-yarn, probably off a bit for medical stuff
- Martin: more pangeo and bugs, some time off
- Matt: Dask bugfix release, track down some intermittent failures in dask/distributed, push dask-jobqueue to releasable state, hopefully some prose writing
- Mike:
- Tom: Dask ordering issues for Dask-ML
- Scott: Finish Hyperband in Dask-ML
- James: Look into adding more scoring metrics for dask-ml

2018-06-14

What we did

- Jim: Lots! Test coverage up for skein 95%, don’t know how to measure coverage in Java. Released conda-pack. Getting surprisingly large feedback. Documentation for skein
- Martin: Started learning pangeo, audit of pickle.loads, more bugs
- **Matt**: Readied things for release. Updated Pangeo deployment. Worked through issue backlog.
- Tom: Pandas / cyberpandas released. More work on gridsearch[incremental[scikit-learn]]
- Mike:

- Scott: Started providing partial_fit support for gridsearch, fleshed out roadmap: https://docs.google.com/document/d/1jsCmPcXIXsSLgdFYgXgngj_P1EkumwZ3MrjkoVaMTjY/edit#heading=h.k37776mpldjz
- James:

Agenda items

- Releasing if there are no objections
- Thoughts on how to humanize the developer and user community?
    - I suspect that more people would contribute if they saw people like themselves contributing.
    - Collect and publish bios, especially from newer contributors?
    - It creates a line on who to invite
    - Other ideas
        - Sprints
        - Tagging issues
        - Better developer docs
    - Blog
        - Would be good to invite people to submit to a blog
        - Maybe copying the content is fine
- SciPy sprints
    - Who will be there? Jim, Scott, Martin until Sunday.
    - Examples sprint
    - Easy contributions
        - Sift through closed issues
        - Sift through Stack Overflow
- Blogposts from Scikit-image/learn/dask meetup
  Nothing has happened yet. Should we start something?

What we plan to do and availability

- Jim: dask-yarn!, release skein 0.0.3. Probably out for a few days for medical stuff
- Martin: ramp up on pangeo, e.g. issue
- Matt: Client meeting in DC next week M/Tu, I would like to do some prose writing otherwise. Releasing and managing fallout.
- Mike:
- Tom: Anaconda webinar (Th), Dask-ML release, off tomorrow
- Scott: Aim to have Hyperband complete by webinar.

2018-06-07

What we did

- Jim: BIDS, skein, started on dask-yarn.
- **Martin**: not much (bought a house!), various bugs; working with Albert DeFusco on examples ideas (tutorial backport is done but not merged)
- Matt: Company meetings
- Tom: Incremental meta-estimator, various dask-ml bugfixes, debug distributed scheduler job, pandas release prep, BIDS sprint wrapup blogpost
- Mike:
- Scott (not present during meeting): started SGD implementation, attended ML conference, cleaned Hyperband and extension nailed down
- James: Helped out with a couple of issues (see #3545, #3560), seeking feedback on how to approach resulting PRs

Agenda items

- DataFrame.sample PR: https://github.com/dask/dask/pull/3566
- Anything more to do on YARN?
- Solved the disk space tutorial issue: https://github.com/dask/dask-tutorial-infrastructure/issues/1

What we plan to do and availability

- Jim: dask-yarn
- Martin:
- Matt:. Working through issue backlog, focusing on distributed scheduler issue
- Mike:
- Tom: Pandas release, Anaconda webinar prep, review the Hyperband PR.
- Scott (not present): Have basic ParamServer/SGD implementation working

2018-05-31

What we did

- Jim: Sprint, Skein is fully functioning, contacted users and am now responding to issues. Hoping to get docs and tests up and a POC of dask-yarn by meeting on Monday.
- Martin: store, render docs, more tutorial backport, distributed timeouts
- Matt:. PubSub
- Tom: Sprint. Benchmarking various things, thinking about delayed fit APIs.
- **Mike**: AE5
- Scott: Integrated dask.distributed into hyperparam alg

Agenda items

- Sprint notes: https://hackmd.io/hN9unCbZTuuEt53kFWUVDQ?edit#

What we plan to do and availability

- Jim: Nicities for Skein, dask-yarn wrapper
- Martin: complete tutorial, move to examples
- Matt:. In California Monday. Austin Visit Tu-W. Back to normal on Thursday.
- Mike: Getting AE 5.2 GPU support production ready. NYC Thurs-Fri.
- Tom: Pandas release Friday. Blogging.
- Scott: Test, benchmark and (maybe) integrate with DaskBaseSearchCV. Will be absent Wednesday and Thursday next week. Today is last day at UW.

2018-05-24

What we did

- Jim:
- Martin: cupy ufuncs, astro discussions, start reworking dask-tutorial, little rechunk
- Matt: PubSub (for some ML applications), trying to make x.T.dot(x) take less memory (auto-rechunk, pausing workers during communication, ordering fixes in distributed), configuration cleanup.

- **Tom**: sprint prep, scalable cross validation, scalable random datasets, (S)GD research, cyberpandas release
- Mike: AE5 stuff, astro discussions, started on JLab / GPU infrastructure
- Scott: Hyperparamer optimization: have an alpha, (dask.distributed not integrated)

Agenda items

- Opportunities for funding astronomy collaborations
- Anaconda involvement in grant proposals for medical imaging
- Dask Release constraints
- FYI, scikit-image/scikit-learn sprint next week

What we plan to do and availability

- Jim:
- Martin: expect clarity on time allotment to Intake next week; meanwhile, working on examples and tutorial repos
- Matt:. More of the same. Flying out today for California. Hopefully also resolving some pangeo deployment issues.
- Mike: Same
- Tom: Travelling to California on Saturday.
- Scott: Finishing up some grad school work, on Anaconda payroll starting Friday

2018-05-17

What we did

- Jim: pycon, skein cleanups and prep for release, engage with potential users
- Martin: pycon; json, zarr, start rechunk as separate PR, start cupy ufunc reading
- **Matt**: pycon, auto-rechunking, anaconda webinar, dask-jobqueue/kubernetes config good to go, finished up dask.set_options -> dask.config.set work
- Tom: PyCon, pandas release, dask release
- Mike:
- Scott:

Agenda items

- Takeaways from pycon
  - Tutorials were useful
  - Chatting with people was useful

    - Might want to avoid having everyone going to the same conference
      But on the other hand it's hard to give a tutorial with one person
    -
- Astronomy uptake
    - Some interest from a variety of parties

What we plan to do and availability

- Jim: release skein, continued work on skein, heading to san jose
- Martin: cupy, rechunk
- Matt:. Rechunking for computation, configuration, recording webinar,
- Mike:
- Tom: Prep for Dask sprint, A couple minor cyberpandas features
- Scott:

2018-05-10

What we did

- Jim:
- **Martin**: to/from_zarr in dask.array, passed on rechunking for the moment
- Matt: pubsub prep, configuration is in, some downstream work
- Tom: tutorial prep, pandas work
- Mike: anaconda enterprise work
- Scott:

Agenda items

-

What we plan to do and availability

- Jim:
- Martin: array rechunking, data catalog for pangeo
- Matt:.
- Mike: busy with other work
- Tom: pandas release, some prep for sklearn sprint
- Scott:

2018-05-03

What we did

- **Jim**: Off a bit for medical stuff, Yarn
- Martin: mostly intake still; intake-xarray example for pangeo, zarr methods for dask.array (thoughts on rechunk), diagnose s3 test failures
- Matt: Configuration, get= -> scheduler=, release
- Tom: Pandas RC. Dask Tutorial prep
- Mike: AE5, intake-xarray
- Scott: Playing with hyper-parameter optimization. Hi! I'm @stsievert on github

Agenda items

- Adding new members
  -
- Release plan
  - New bugfix release for s3/arrow?
  - Merge configuration, scheduler=, other biggish changes
- Pycon tutorial prep
  - Cloud credits (Ben is checking)
  - Demo Friday tomorrow to stress test a bit
- What is the status on the filesystem work?
  - Community improvements coming in s3fs and hdfs3
  - Is anything blocking progress?
- What is the status on Yarn work?
  - CLI works for starting, getting status, and stopping jobs
  - Can successfully run dask
- MLCON Dask slide?

What we plan to do and availability

- Jim: Yarn. Tutorial. Prep for tensor decomposition work
- Martin: tutorial
- Matt: Finish off current largish issues. Maybe some kubernetes/pangeo work.
- Mike: AE 5, intake-xarray (night gig)

- Tom: Pandas release. PyCon tutorial prep and tutorial
- Scott: keep playing with hyperparam opt

2018-04-26

What we did

- Jim: A bit of yarn work, off for medical stuff
- Martin: make file system spec for comments ; handed off serialization
- Matt: Vacation, also configuration, serialization, and small issues
- **Tom**: Pandas
- Mike: AE5

Agenda items

- Update on using dask in other scipy tutorials
    - https://github.com/dask/scipy-tutorials-2018
    - Reached out to authors in https://github.com/dask/scipy-tutorials-2018/issues/3

      So far no unexpected takers.
      Joris (geopandas) and sklearn developers (other than Andy) have mentioned that they'd be game though. Maybe we have to reach out group-by-group?
- Pycon tutorial prep
    - Cloud credits (Ben is checking)
- What is the status on the filesystem work?
    - ...
    - Is anything blocking progress?
- What is the status on Yarn work?
    - Trying to get two ends to talk to each other
    - Maybe a demo friday demo
- Nbserverproxy & AE5
    - Depends a bit on how dask is being deployed. Will it be on the same kubernetes cluster as AE5 or a separate one.
    - Talk to Ben.

What we plan to do and availability

- Jim: Yarn. Research for tensor decomposition
- Martin: intake mostly ; xarray plugins for pangeo
- Matt:Back full time next week, not sure what topics in particular
- Mike: AE 5, intake-xarray (night gig)
- Tom: Pandas release. PyCon tutorial prep.

2018-04-19

What we did

- Jim: Yarn work, a few user issues, meetings
- Martin: mostly intake; some efforts at deserialisation in distributed #1912, gcsfs instance reuse and more credentials fun
- **Matt:** Talks at Inria and ICDE, some issue triage
- Tom: Vacation
- Mike: Dreaming of intake, xarray, EarthML. Still focused on AE5. Will have a transition timeline soon!

Agenda items

- Update on using dask in other scipy tutorials
  - https://github.com/dask/scipy-tutorials-2018
  - Informally sklearn, geopandas tutorial leaders think that it's a nice idea
  - Some discussion around exactly how to do this on the JupyterHub / software environment side
  - Will probably reach out to authors soon
- Issue on generic ndarray algorithms for tensorly https://github.com/tensorly/tensorly/issues/47
- Dask datasets: https://github.com/dask/dask/issues/3397
  - Tutorials have some content here
  - We can also have data in s3, do we want to depend on this?

    - Intake?
    - We should think about consistency across datasets
    - Files or direct dask objects?
- Status of yarn work
    - https://github.com/jcrist/skein
    - Pain with unfamiliarity of java issues. Mike has some familiarity and can help.
- Long term maintenance of gcsfs, s3fs, fastparquet, …
    - Gcsfs: google seems uninterested in maintenance so far
    - S3fs and fastparquet: not much time needed

What we plan to do and availability

- Jim: Yarn. Research for tensor decomposition
- Martin: intake mostly (get things ready for testing), serialisation, filesystems maybe
- Matt: Vacation half-time,
- Mike: AE 5?
- Tom: Vacation

2018-04-12

What we did

- Jim: Time off for family, some time at anacondacon, some Yarn stuff, things are talking to each other
- **Martin**: annoyance with moto, a little streamz fun, continuing gcsfs auth issues on clusters
- Matt: Wrote some slides, small bugfixes in distributed, anacondacon
- Tom: AnacondaCon, pandas dev meeting with release as-is
- Mike: AnacondaCON, AE5

Agenda items

- Results of pow-wow
    - 1.0
- Scipy talks

What we plan to do and availability

- Jim: Yarn. Research for tensor decomposition
- Martin: secure client, non-pickling
- Matt: Issues, Paris next week
- Mike: AE 5
- Tom: Vacation. Maybe pandas at the end.

2018-04-05

What we did

- **Jim**: Mostly off, family is in town. Some yarn work, now (kind of) able to start jobs in python alone.
- Martin: mostly intake since return from workshop; fastparquet release; some streamz cleanup; orc finally in
- Matt: Dask.distributed cleanup, particularly around starting workers and memory leaks. Looking into serialization a bit.
- Tom: Pandas / cyberpandas. Anacondacon prep
- Mike:

Agenda items

- Discussion on big-data caching
    - Motivated by conversations with HEP groups
    -
- Aiming to release dask.distributed tomorrow. Any objections?
- Suggestions on members to add to Dask github org
    - Will raise as an issue

What we plan to do and availability

- Jim: Yarn Work, starting research on tensor factorization, AnacondaCon???
- Martin: bit behind on intake, catching up on streamz and dask issues
- Matt: Real-time dask talk at AnacondaCon and prep for that. Release. Some client work around avoiding pickle deserialization. Out at conferences/vacation the next three weeks
- Mike:

- Tom: Travelling to Austin on Friday / Saturday. Anacondacon next week. Vacation the week after. Working to get a pandas release by the end of April.

2018-03-29

What we did

- Jim: Yarn work. Been building individual components to ensure my understanding of things was correct, working on uniting them over the next week.
- Martin:
- **Matt:** Dask.distributed issue triage. There have been some issues with starting up workers in slow environments. Documentation.
- Tom: pandas extension work is maybe wrapping up
- Mike:

Agenda items

-

What we plan to do and availability

- Jim: continue work on yarn, working half time - family's in town
- Martin:
- Matt: AnacondaCon prep. Documentation and blogposts. Distributed worker startup failures.
- Mike:
- Tom: dask.dataframe docs, dask-ml issues, anacondacon prep

2018-03-22

What we did

- **Jim**: yarn work. Mostly experimenting so far, started writing actual code yesterday.
- Martin: made talk, more kafka, mostly Intake (new: hbase, hdfs secure)
- Matt: handled backlog of issues (tornado, task ordering) from previous week, release, docs, starting to push on joblib a bit
- Tom: pandas / cyberpandas. Some dask-ml performance investigation.
- Mike:

Agenda items

- Binder examples (matt)
  Good opportuntiess for early contributors
  Need to have clear contribution guidelines
  Probably good for scipy sprints

What we plan to do and availability

- Jim: continue work on yarn
- Martin: trip to HEP conference
- Matt: docs, some distributed worker failures, anacondacon prep. Less responsive starting early April
- Mike:
- Tom: pandas / cyberpandas

2018-03-15

What we did

- Jim: Dockerized hadoop infrastructure, started on yarn work
- Martin: orc interface to arrow's orc reader (written by Jim); kafka streamz improvements; start HEP talk

- Matt:
- **Tom**: Cyberpandas / pandas. A bit of dask-ml benchmarking.
- Mike: NASA - fixed unit tests for earthio, looked over sklearn xarray libraries, planning path forward. Shipped AE 5.1.1

Agenda items

- Scott Sievert interning for Anaconda

What we plan to do and availability

- Jim: Continue on yarn work
- Martin: make that talk! Try to look into tutorial reported bugs
- Matt:
- Mike: NASA notebook slunking; xarray -> sklearn -> xarray paths.
- Tom: pandas / cyberpandas.

2018-03-08

What we did

- Jim: sick all week
- **Martin**: (time off for teeth); restart work on kafka in streamz
- Matt: Adaptive deployments, cluster superclass, tornado issues,
- Tom: Strata
- Mike: AE Platform and spinning up on ELM

Agenda items

- Github membership (matt)
- Documentation mainpage
- Martin work
    - Streamz + kafka
        - Asynchronous issues
        - Distributed reading with kafka
        -
    - Orc reader + Dask

        - Lots of changes in the future, but they'll probably be for newer features. Should be fine for pulling chunks of rows/cols

What we plan to do and availability

- Jim: Not sure, depends how well I'm feeling. Hopefully yarn work.
- Martin: look into orc...
- Matt: Release, try to settle streamz asynchronous issues, out next week on vacation
- Tom: cyberpandas

2018-03-01

What we did

- **Jim**: Client work, CI infrastructure for yarn work
- Martin: gcsfs release, parquet issues (again)
- Matt: Docs, kubernetes-helm (almost in!),
- Tom: Pandas, a bit of dask-ml docs

Agenda items

- Github Membership (Matt)
- Pycon arrangements
    - Tom will book Tuesday - Sunday
- Core things for Martin to do in Dask
    - Streaming
    - Orc reader - not quite ready on the single-machine side
    - GeoPandas -
        - Maybe Joris will handle things.
        - This may not be expensive for someone familiar with the project
    - Merge PRs in libhdfs3 ?
        - Hoping that this becomes unnecessary if other projects take over
    - Avro-arrow reader (old PR by Marius)
        - There are a few options here, uavro, cyavro, arrow-avro
    - Arrow for inter-worker pandas serialization
    - Hive metastore (also old PR by Marius)

What we plan to do and availability

- Jim: Yarn work
- Martin: some time off for dentistry
- Matt: Finish Kubernetes work (hopefully), documentation, On vacation March 13th-18th
- Tom: Strata San Jose next week

2018-02-22

What we did

- Jim: Scipy tutorial proposal, not much else, working on other projects
- Martin: httpfs, some polish to memoryfs (if anyone wants that), speed up fastparquet.dataframe.empty, merges on gcsfs
- Matt: Documentation, both landing page and walking through normal docs. Cleaning up some user issues that had accumulated.
- Tom: Cyberpandas / pandas. Webinar prep.

Agenda items

- Micro-release (Matt, estimated 7 minutes)
- Pycon arrangements

What we plan to do and availability

- Jim: Yarn work
- Martin: plan for HSF workshop; more gcs?
- Matt: Release. Go through docs.
- Tom: webinar prep, proposals, pandas / cyberpandas

2018-02-15

What we did

- Jim: Finished a few lingering issues last week, time off all this week
- Martin: (time off sick) No great progress on gcsfs. HTTP file system
- Matt: Release. Docker images and new helm chart. Added Bokeh plots for node-link task graph
- **Tom**: dask-ml release. Cyberpandas / pandas things

Agenda items

- SciPy talks (feb 15 deadline for tutorial/talks)?
    - Tutorial
        - Jim will do a dask data analysis tutorial today
        - Coordinate with other tutorials to provide distributed resources (scikit-image, xarray, scikit-learn)
    - Talks
        - Jim- maybe approximate algorithms
        - Matt- daskboard
        - Martin + Mike McCarty - astro
        - Reach out to sklearn folks for Dask-ML talk
- Change meeting time?
  Change video conferencing system?

What we plan to do and availability

- Jim: SciPy proposals, start on the yarn work
- Martin:
- Matt: Webinar prep. Write about kubernetes. Maybe some more bokeh things.
- Tom: webinar prep, proposals, pandas / cyberpandas

2018-02-08

What we did

- Jim: Some arrow work, user issues, debugging dask memory usage issues with a client, rebooted work on crick
- Martin: Quite some time investigating gcsfuse improvements: found some improvements both on directory listing (thanks @asford) and chunk caching; needed to fix readahead behaviour
- **Matt**: Closed out dask.distributed work, ready for release, testing and cleaning up Daskernetes, starting work on our docker stack.
- Tom: Pandas. dask-ml meta estimator

Agenda items

- SciPy talks (feb 15 deadline for tutorial/talks)?
    - Tutorials
        - Parallel tutorial
            - Ben, Min, Olivier, Aron, Tom
            - Matt to send out e-mail
        - Dask tutorial
            - Jim wrangling others
        - Maybe include dask-ml work in sklearn tutorial?
            - Tom will e-mail people
    - Talks
        - John Kirkham may submit
        - Astro and Dask talk
            - Mike McCarty
            - Martin
        - Dask-ML talk
            - Maybe co-present with sklearn folks
        - Who else can we encourage to give talks?
            - Pangeo
            - Sparse arrays
            - …?
- How easy it is to write an Intake plugin that dask can make use of
- Release blockers
    - https://github.com/dask/dask/pull/3132
- Yarn

    - Martin has a start on a docker testing setup: https://github.com/martindurant/docker_images/tree/master/kerb-hadoop
- PyCon:
    - Who's going? Tom, Martin, Jim, Matt

What we plan to do and availability

- Jim: Out on vacation thru tuesday, SciPy proposal(s), start yarn work, probably user issues
- Martin: more tests on gcsfuse; scipy proposal(s)
- Matt: Release. Release and document daskernetes. Clean up helm charts. Webinar prep?
- Tom: Pandas (probably for a couple weeks at least). Maybe a dask-ml release.

2018-02-01

What we did

- Jim: PyArrow support for HDFS, closed stale issues, bytes refactor, a few user issues
- **Martin**: released s3fs, fastparquet, python-snappy (pypi only); fixed fastparquet regression and sql dtypes
- Matt: testing issues in dask.distributed, task prioritization in dask and dask/distributed, some small fixes prior to release. Sparse docs:http://sparse.pydata.org/en/latest/ , some blogging.
- Tom: Visited Inria, various joblib, distributed, dask-ml improvements. Started blog posts recapping. Prep for AnacondaCON.

Agenda items

- SciPy talks (feb 9 deadline for tutorial/talks)?
    - Tom is reaching out to others about presenting (Skipper, Olivier)
    - John Kirkham may submit
    - Dask-ML: Tom is interested in doing a talk, but may not be around
- PyCon Tutorial Prep?
    - Verify that we have credits.
    - See if Ben has any interest in helping.
- Report from Inria visit?
- dask-image subpackage

What we plan to do and availability

- Jim: A few arrow and dask issues needed for client work, perhaps start on yarn work, perhaps cleanup crick for inclusion into dask for approx methods
- Martin: review arrow/hdfs, small fastparquet fixes
- Matt: Close current round of dask.distributed PRs. Release. Probably return to Kubernetes afterwards.
- Tom: Mostly cyberpandas things.

2018-01-25

What we did

- Antoine:
- Jim: dask.bytes, hdfs cleanups, adding pyarrow hdfs support in PR (exposed many bugs)
- Martin: made fixes to fastparquet and sql; gcsfs *finally* released; answering issues
- Matt: Clean up task ordering (waiting on passable tests). Released sparse. A variety of conversations and support around Pangeo work and grant writing
- Tom:
- Erik:

Agenda items

- Status update on testing failures
    - Fastparquet issues maybe fixed? Not sure on dev pandas
    - Dev builds moved to only test on merge and allow failures
    -
- Documentation page linking to external projects
    - Must have doc page, CI testing, installable, responsive to issues/PRs
- Style guide
    - Yapf: flake +
    - Perhaps just https://google.github.io/styleguide/pyguide.html is sufficient? I (Jim) agree with most of this, and dask for the most part follows these guidlines
    - Con:

        - Maintaining a style guide can be expensive
        - The scope is very large
    - Pro:
        - It’s nice to have conversations once
        - We don’t have to write down everything at once, it can be an evolving document
- SciPy
    - Deadline is February 9th
    - Tutorial?
        - Same as last year?
    - Talk?
        - No concrete proposals at the moment, lets do some thinking
- Gcsfs authentication
    - Auth issues may be resolved with new google-auth library
    - UK Met folks mentioned that it was valuable to mount all of /s3 or all of /gcs

What we plan to do and availability

- Antoine:
- Jim: Finish hdfs issues, hopefully have time to get back to lingering PRs
- Martin:more work on intake this week; maybe fastparquet release; looking into file buffering for fuse
- Matt: Around all week. I would like to get dask/distributed into a releasable state. Tending pangeo.
- Tom:
- Erik:

2018-01-18

What we did

- Antoine:
- Jim: more `dask.bytes` work, a few user issues. Out with allergies for a bit :(
- Martin: gcsfs released, fastparquet issues
- Matt: task prioritization in dask.order. Some administrative work on pangeo. Anaconda Inc. company meetings.
- Tom: More cyberpandas
- Erik:

Agenda items

-

What we plan to do and availability

- Antoine:
- Jim: More `dask.bytes` work.
- Martin: iron out as many quick fastparquet issues as possible and release
- Matt:
- Tom: Visiting Inria next week
- Erik:

2018-01-11

What we did

- Antoine:
- Jim:
    - A few dask user issues
    - Refactoring dask.bytes work
    - Some pyarrow stuff
- Martin: Revamp auth in gcsfs, finally getting fuse working (testing in pangeo)
- Matt:
- Tom: Almost entirely on “cyber pandas”
- Erik:

Agenda items

- Dask and intake
-

What we plan to do and availability

- Antoine:
- Jim:
    - Continue refactor of dask.bytes

    - Start on filesystem spec work, share filesystems between arrow and dask
- Martin:
    - Finish off gcsfs, release?
- Matt:
- Tom: More cyber pandas, prep for visit to sklearn dev
- Erik:

2018-01-04

What we did

- Antoine:
- Jim:
    - Vacation
    - Fixed a few dask issues
    - Finished ORC reader
- Martin:
    - GCS: Fuse for reading HDF5, and other related issues around connection persistence and caching
    -
- Matt: Working on basic pangeo deployment.
    - Dynamic Kubernetes deployment: https://github.com/yuvipanda/daskernetes
    - Some admin work on pangeo.pydata.org (currently offline)
    - Futzing with JupyterHub
- Tom: Mostly pandas / client work on extending pandas with custom arrays
- Erik:

Agenda items

- Release (6 minutes, Matt)
-

What we plan to do and availability

- Antoine:
- Jim:
    - Currently working on a few dask issues that are blocking work
        - Use hdfs from arrow instead of hdfs3
        - Weird mildly non-reproducible deadlocking behavior in our custom graph

    - Fix **Issue #2885: `dask.bytes.core.get_fs_paths_myopen` should strip protocol/options from paths** by jcrist in dask/dask on GitHub
    - Filesystem spec library
- Martin:
    - More gcsfs
- Matt: In Austin this week. At conference Monday/Tuesday. Plan to prepare for and then deliver upcoming talks.
- Tom: pandas-ip, O'Reilly article
- Erik:

2017-12-21

What we did

- Antoine: sick for several days; did a bit more on cythonization, results still mediocre
- Jim: Respond to feedback on ORC PR, non-dask work
- Martin
- Matt - Was at AGU last week. Working on documentation, PR backlog, and experimenting with Kubernetes/Jupyterhub deployments since then. Change dask.order logic. https://github.com/dask/dask/pull/3017
- Tom: pyarrow 0.8 parquet, pandas things

Agenda items

- Meeting logistics (Matt, 7 minutes)
- Cythonized scheduler

What we plan to do and availability

- Antoine: available starting Tuesday (27th) up to Thursday
- Jim: Probably mostly off
- Martin
- Matt - working half time
- Tom: Off Friday - ~Wednesday. Working on pandas things otherwise
- Erik: client work, has some free time coming up. Send issues his way