Motivation

Today Flink runner uses `DataStream` API for streaming jobs and `DataSet` API for batch jobs. Flink `DataSet` API is in a semi-deprecated status and will be removed in some future major version. Flink `DataStream` API has been extended to support batch jobs as well. Therefore, it makes sense to completely remove `DataSet` API usage from the Flink runner and converge the batch and streaming execution path to `DataStream` API.

Migration Plan and Public Interface Change

- To maintain backwards compatibility, a new configuration of `use.datastream.for.batch` will be introduced.
    - If this configuration is set to `TRUE`, the Flink runner will use `DataStream` API to run the batch jobs. Otherwise, the batch job will be run with the existing `DataSet` based implementation.
- The default value of `use.datastream.for.batch` will be `FALSE` for the existing Flink versions.
- When new Flink version (e.g. Flink 1.16) is added, we will set the `use.datastream.for.batch` to `TRUE`.
- After 6 months, we will mark the `DataSet` based support as deprecated. And users who set `use.datastream.for.batch` to `FALSE` will receive a warning.
- After 18 months (or when Flink community removes the code of `DataSet`), we will remove the `DataSet` related code completely.

Proposed Changes

I have implemented a prototype which migrates the batch jobs to use Flink `DataStream` API. The POC has passed all the existing unit tests and runner validation tests. The POC only implemented the migration on the classic Flink runner. The portable runner should be very similar.

It looks like there isn't a significant amount of work for the `DataSet` to `DataStream` migration. Some key changes are:

1. Use FLIP-27 Source instead of SourceFunction / InputFormat.
2. Fix some PTransform URN differences between the `FlinkBatchTransformTranslators` and `FlinkStreamingTransformTranslators`
3. If `use.datastream.for.batch` is set to `TRUE`, let `FlinkStreamingTransformTranslators` handle both streaming and batch job translations.

4. Modify `DoFnOperator` to make it support both stream and batch mode.