Requirements for the Panda DB setup for sPHENIX at SDCC

- **Responsibilities of each party**
  - NPPS
    - Full DB schema management;
    - DB content management;
    - Migration of the data from the master DB to the archive DB if there is going to be one;
  - SDCC
    - DB hardware management, VM management, Storage management;
    - Continuous DB monitoring (availability, number of clients per VM, space usage, **more?**);
    - DB backups and recovery as specified by policies defined by sPHENIX and NPPS teams;
    - DB server software upgrades required to stay compliant with the BNL Cyber Security policies;
    - Access control management and scheduled password updates as per BNL Cyber Security policy;

- **Scalability and Usage**
  - What is the expected total number of clients / concurrent jobs ? 70k…150k?
    - **Answer: Maximum ~270k (2025) #cores**
  - How many input/output datasets/files per job, task?
    - **Answer: (min, max, avg)**
      - **Reconstruction: 61 inputs, 2/3 files output**
      - **Tracking: 41 inputs, 2/3 outputs**
      - **Calorimeter: 21 inputs, 2/3 outputs**
  - What is the expected number of simultaneous/persistently-connected clients and/or queries per second?
    - **Answer: Maximum: writes ~1k / sec, reads ~15k / sec (CERN)**
  - What kind of load is expected? Would it be mostly reads, or an equal mix of reads and writes, or mostly writes?
    - **Answer: ~10:1 reads:writes**
  - ~~Does Panda use an ORM / DBAL?~~ It is understood that Panda is primarily used with Oracle DB. What about PostgreSQL or MySQL? Is the Panda-PostgreSQL setup at BNL well-tested and considered reliable?
    - **Answer: PanDA supports both Oracle and PostgreSQL. DOMA PanDA@CERN is in production with PostgreSQL (version >= 14)**
    - **PanDA supports two accounts panda_read, panda_write**

- ~~Does Panda support internal load balancing of any kind? I.e. does it have the ability to balance/spread its own load between multiple DB servers natively? To distinguish between read and write operations?~~
    - PgPool-II on replicas will split writes from reads
  - Is there going to be a dedicated “archive” DB server (e.g. panda job archived)
    - **Answer: YES**
    - , or all three years of sPHENIX running are expected to be covered by the master DB?

- **Technical Specs**
  - What is the recommended hardware configuration for the Master DB (=writes) in terms of CPUs, RAM, Disk space, and Networking?
  - ~~What is the recommended hardware configuration for the DB Replicas (=reads, if load balancing is supported)?~~
  - What is the required DB Storage size for the expected period of running?
    - **Answer: ~30TB**
  - What is the recommended DB service version (assuming PostgreSQL/MySQL)?
    - **Answer: PostgreSQL ver 14 or higher**
  - Are there any specific requirements for the network speed ?
    - **Answer: No specific requirements, but the system is OK when DB and PanDA server VMs are in Meryin network at CERN, while it experienced latency issues when some VMs were in Wigner.**
  - Are there any specific requirements for the network layout (isolation to specific subnetworks)? Currently, all sPHENIX jobs are expected to be run on SDCC nodes only, so if there is a need for external access, then network conduits should be specified and requested.
    - **Answer: It would be nice if pandamon:443 is reachable outside of SDCC/BNL network, so that non-BNL-based people can see jobs/tasks without going through VPN.**
      - **-> Requirements to discuss with Chris Hollowel**
      - **Outside BNL connectivity**

- **Backups and Disaster recovery**
  - What are the desired backup frequency and retention time? Ex.: “backups are expected to be taken daily, retention time two weeks, + additional monthly/weekly backup with the retention time of one year”
    - Full backups are preferable, yet a disk space estimate is needed to calculate the backup storage space
    - **Answer: Better to check with CERN DBAs, but it is supposed to be daily backup.**
  - Is there a Panda-specific recommendation for the High Availability DB setup
    - **Answer: No**

    - What is the expected time to restore the database from backup in case of any disaster event that caused the DB downtime? Is this service going to be a mission-critical service that will need 24/7 operator shifts?
        - **Answer: Depends on sPHENIX requirements.**
            - 24/7 support during run for many services hpss, rucio, okd, pandaDB, etc
    - Are there going to be regular DB maintenance periods coordinated with the sPHENIX online data-taking team?
        - **Answer:**
            - uptime 100% during runs vs security patches (should be fine on the sdcc intranet)

- **Possible DB deployment configurations**
    - ~~**1 Master DB + 1 Hot Standby Replica** (continuous replication) - basic High Availability setup (HA), Replica is not used by Panda, and is serving as a "hot swap" for master DB. Also, Replica is used for backup purposes;~~
    - **1 Master DB + 3 Hot Standby Replicas** (continuous replication) - assumes the possibility to split write/read operations between Master and Replicas (=PgPool-II splitter). Provides better HA. Two Replicas are used to balance the read operations of clients, one Replica is used solely for backup purposes (avoids backup overhead on Master/active Replicas);
    - ~~**DB Clustering** solution? (for PostgreSQL, it likely means purchasing the proprietary software license). Clustering allows even better HA and allows sustaining the high volume of write operations;~~

Timeline:

- Runs in April
- As soon a possible

Contact channel:

- RT queue
- Mattermost

Tech. contact : Xin