From Oracle to Parquet and Spark: Enabling Scalable Access and Analysis of ATLAS DCS Data

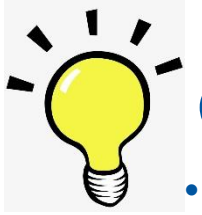
ATLAS Scope meeting
February 12th, 2026
Luca Canali, ADAM and ATLAS-DBA

Luca Canali – A Brief Intro

- DBA and Data Engineer
- I am part of the ATLAS DBA team
  - I work on ATLAS Oracle DBs, helping ATLAS developers and database users, under ADAM coordination
  - I work on helping ATLAS users with using Spark, Hadoop, SWAN

Analysis of DCS Data

- DCS data mostly used for online operations
  - Optimized for writes
  - We only read small chunks of data, for monitoring, etc



Can we get insights by analyzing large chunks of DCS data?

- Working with detector experts for the analyses

Toolset

- Let's not use the DB (Oracle) for this (for stability of online operations, cost, performance)
- Solution: import data into a separate and optimized platform for data analysis

DCS data in Oracle

Offloading Data from Oracle for Analytics

- Key idea: moving data from databases to platforms for analytics
  - Export from DBs into files (Parquet format)
  - Query data with scalable engines (Apache Spark)
  - Worked with RUCIO, EventIndex, ATLAS DCS and NSW

Oracle DB

From Oracle tables to Parquet files

Compute and storage on Hadoop and/or Cloud

Python jobs and/or notebooks

DCS Schemas Offloaded to Parquet

- Data is copied from Oracle to Parquet files
  - Files hosted in a Hadoop cluster provided by CERN IT (a general-purpose cluster called Analytix)
  - Whole PVSS schemas exported one-off
    - In addition, scheduled daily and incremental import for main table (eventhistory table)
    - Detectors/DB schemas currently exported: PVSSMDT, PVSSMMG, PVSSMUO, PVSSRPC, PVSSSTG, PVSSTGC
  - Eventhistory table is partitioned by timestamp values (year, month, day)
    - This provides additional performance boost for selecting data of interest by timestamp range

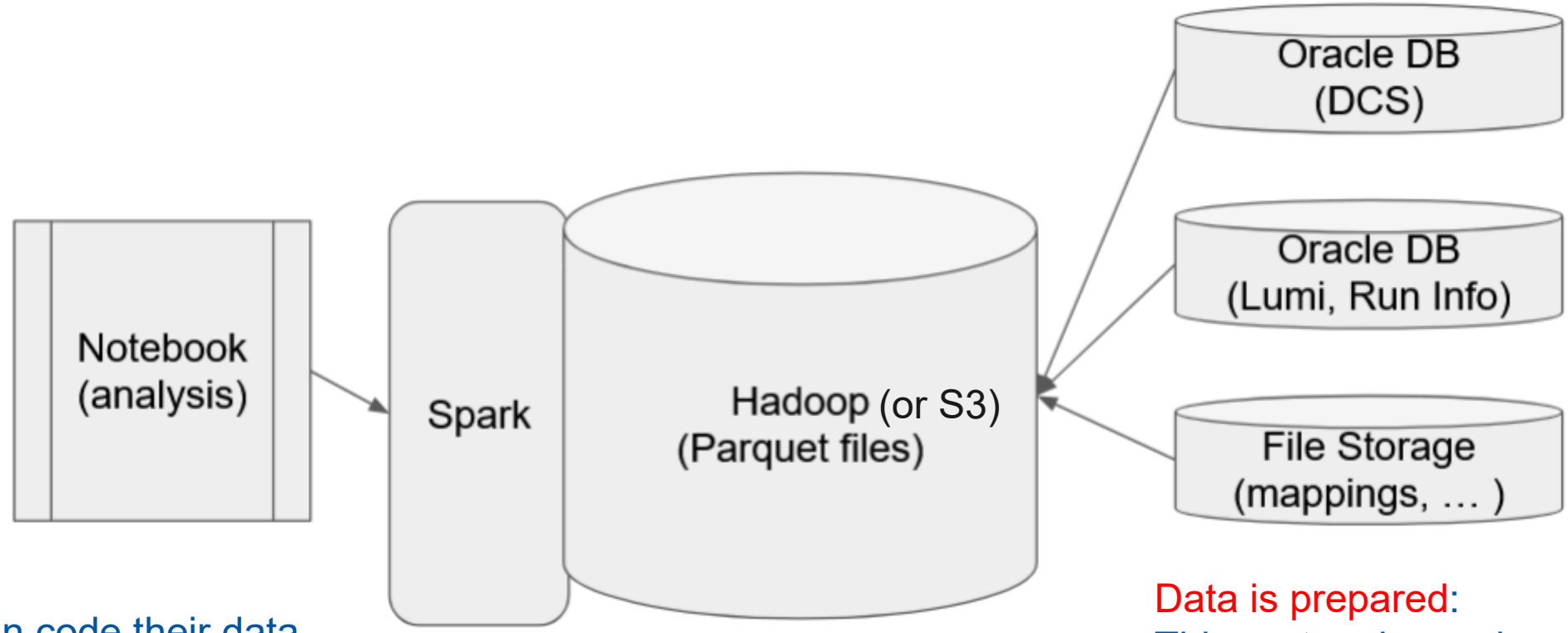
Technology: why Apache Parquet?

- With Apache Spark plus Parquet files we are building a DB for large scale analysis
  - High adoption in industry and open source for building data lakes and data warehouses
- Parquet
  - Is a columnar data format
  - Optimized for storing and querying data for large-scale analysis
  - Uses encoding and compression
  - Data is stored together with its schema
  - Works well with Apache Spark, Pandas, and many other tools
  - Provides a simple way to map DB tables into files

IT Services: SWAN, Spark, and Hadoop

- Data analysis platform, interactive and at scale
  - Running on many CPUs like batch, but interactive (notebooks)
- Storage and compute (Hadoop, Spark) from CERN IT
  - DCS data, used so far (Feb 2026) ~ 4 TB
  - Fraction of a shared cluster capacity: 1500 cores, 21 PB
- SWAN, a CERN service for Python notebooks
  - https://swan.web.cern.ch/swan/
  - Integrates LCG releases and Spark
  - Easy to get started, build from examples

End-to-End Data Analysis Platform



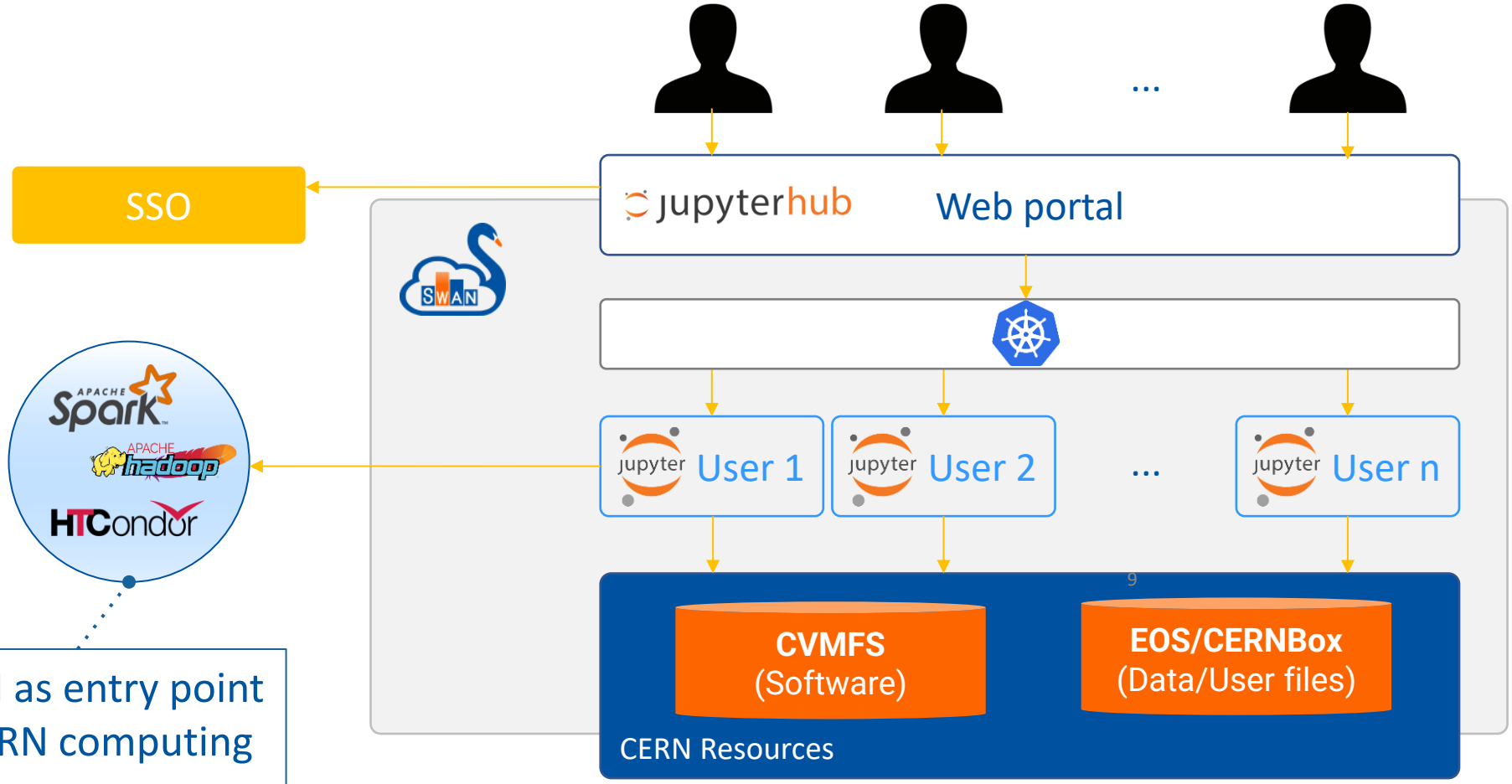
Users can code their data analysis in Python using scripts or Jupyter Notebooks (CERN SWAN)

Clusters are available for execution at scale
Data APIs that work at scale: Data Frames and SQL

Data is prepared:
This system is used as a dynamic database, importing needed information from external data sources into parquet files

CERN SWAN architecture

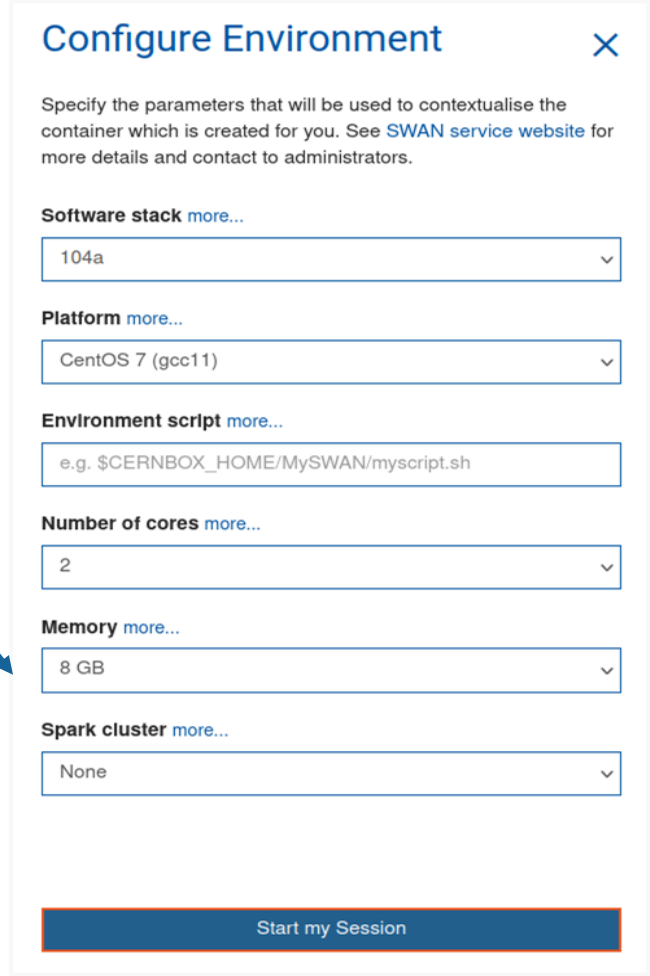
https://swan.web.cern.ch/swan/



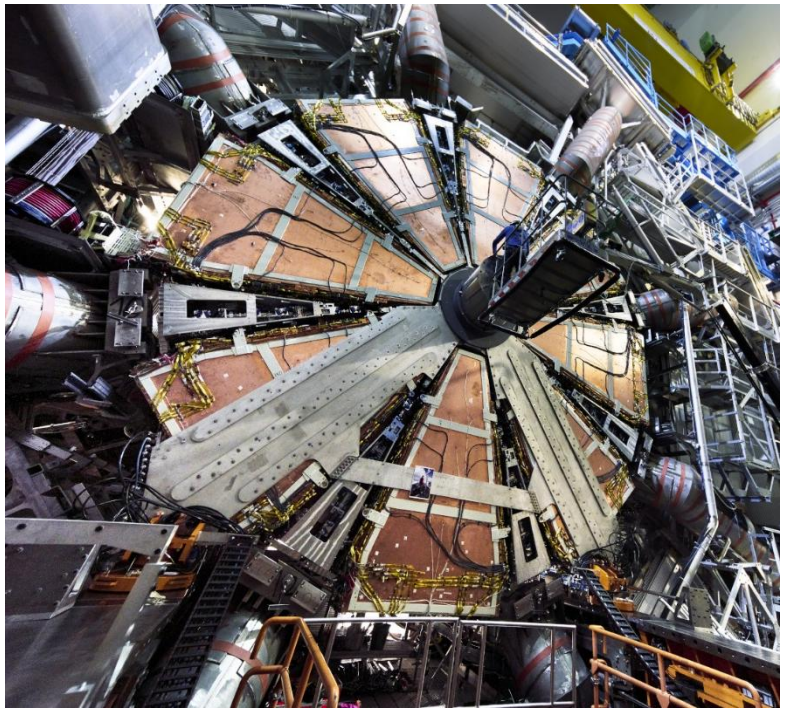
SWAN as entry point to CERN computing resources

SWAN integration with Hadoop / Spark

- SWAN is connected to Hadoop & Spark clusters at CERN
  - Jupyter extensions available to connect to clusters and spawn Spark executors
  - Monitor the execution of Spark jobs
  - Training material on Spark and SWAN: https://sparktraining.web.cern.ch/



Example of Use by ATLAS Community



- New Small Wheel
  - Interested in analyzing DCS data imported into Parquet and joining it with COOL data among others
  - A few NSW experts “onboarded” on the use of SWAN and data analytics tools

- Work by ATLAS ADAM + Michelle Ann Solis (NSW)
  - NSW DAQ Link Stability Investigation in 2024
  - Actions: looking at several parameters monitored in the DCS to help explain instability

- Details:
  - **Advancing ATLAS Detector Control System Data Analysis with a Modern Data Platform** EPJ Web of Conferences **337**, 01110 (2025)

Wrap Up

- DCS data from Oracle to a platform for analysis
  - Pipelines are up and running, maintained by ATLAS DBA team using CERN IT infrastructure
- Analysis on SWAN using Python notebooks
  - Expressive APIs, run at scale on clusters
- Example
  - NSW task force on DAQ link stability investigations
- We are happy to help detector experts to start their analysis of DCS data

- Contact: atlas-dba@cern.ch

References

- Article:
  - **Advancing ATLAS Detector Control System Data Analysis with a Modern Data Platform** EPJ Web of Conferences **337**, 01110 (2025)
- Gitlab project:
  - https://gitlab.cern.ch/atlas-dba/dcs-offload
- SWAN service:
  - https://swan.web.cern.ch/swan/
- Hadoop service:
  - https://hadoop-user-guide.web.cern.ch/
- Training material on Spark and SWAN
  - https://sparktraining.web.cern.ch/

- Acknowledgements:
  - ATLAS DBA and ADAM, Andrea Formica, Michelle Ann Solis (NSW), IT services for SWAN and Spark/Hadoop