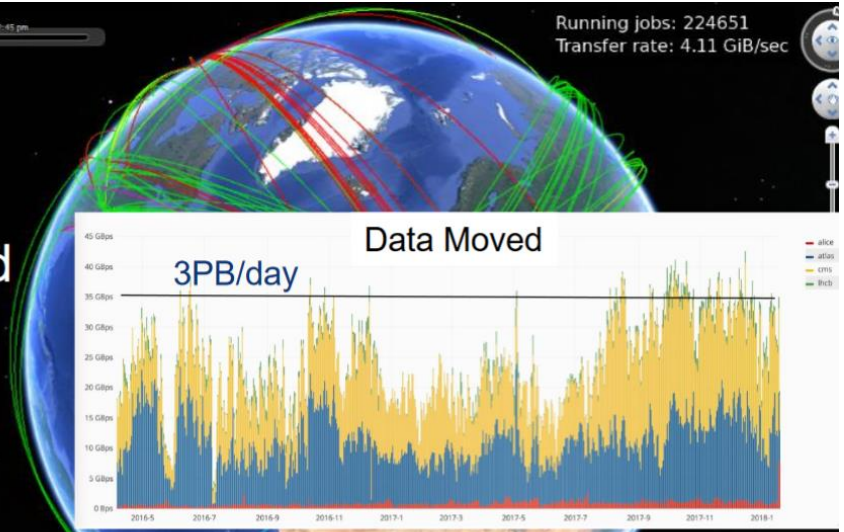
Intro to “Big Data” and Analytics Services

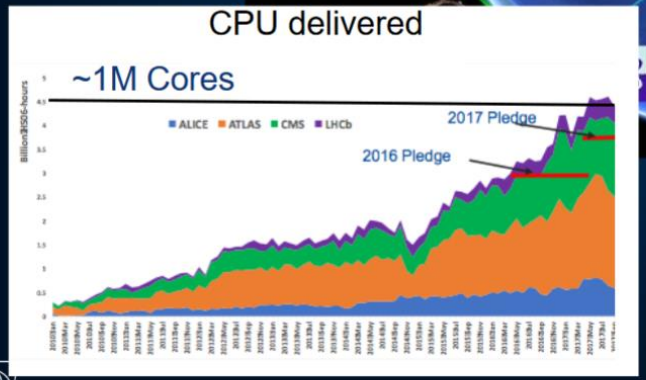
Luca Canali, IT-DB

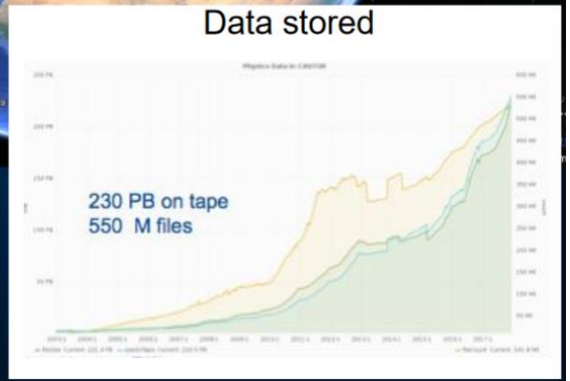
Visit JP Morgan, May 7th, 2018

LHC Data

- Worldwide distribution and processing of LHC data







LHC data processing has custom built solutions and very large scale

Hadoop Clusters at CERN IT

- Several orders of magnitude below LHC data processing systems
- 3 current production Hadoop clusters
  - + environments for NXCALS DEV and HadoopQA
  - Just commissioned a new system for BE NXCALs (accelerator logging) platform
- Numbers relate to the size of the infrastructure (updated Q2 2018):
  - 14 PB Storage, 110 nodes, 3100 logical cores, 20 TB memory

Analytics Pipelines – Use Cases

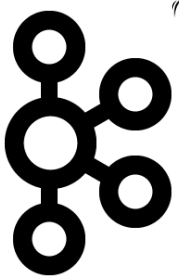
- Many use cases at CERN for analytics
  - Data analysis, dashboards, plots, joining and aggregating multiple data, libraries for specialized processing, machine learning, …
- Communities
  - Physics:
    - Analysis on computing metadata (e.g. studies of popularity, grid jobs, etc) (CMS, ATLAS)
    - Development of new ways to process ROOT data, e.g.: data reduction and analysis with Spark-ROOT by CMS Bigdata project, also TOTEM working on this
  - IT:
    - Analytics on IT monitoring data
    - Computer security
  - BE:
    - NX CALS – next generation accelerator logging platform
    - BE controls data and analytics
  - More:
    - Many tools provided in our platforms are popular and readily available, likely to attract new projects, notably the analytics platform with hosted notebooks SWAN_Spark
    - E.g. Starting investigations on data pipelines for IoT (Internet of Things)

“Big Data”: Not Only Analytics

- Data analytics is a key use case for the platforms
- Scalable workloads and parallel computing
  - Example work on data reduction (CMS Big Data project) and parallel processing of ROOT data (EP-SFT)
- Database-type workload also important
  - Use Big Data tools instead of RDBMS
  - Examples: NXCALS, ATLAS EventIndex, explorations on WINCC/PVSS next generation
- Data pipelines and streaming
  - See example of monitoring and Computer security (Kafka development with help of CM)
  - Also current investigations on IoT (project with CS)

Highlights of “Big Data” Components

- Apache Hadoop clusters with YARN and HDFS
  - Also HBase, Impala, Hive,…
- Apache Spark for analytics
  - Apache Kafka for streaming
- Data: Parquet, JSON, ROOT
- UI: Notebooks/ SWAN



Challenges

- Platforms
  - Provide evolution for HW (Hadoop platform) and SW (distribute and update software and configuration)
- Service
  - Build robust service for critical platform (NXCALS and more) using custom-integrated open source software solutions in constant evolution
  - Support production services (IT monitoring, Security, ATLAS EventIndex)
  - Evolve service configuration and procedures to fulfil users needs
  - Further grow value for community and projects -> SWAN and analytics platform
- Knowledge and experience
  - Technology keeps evolving, need to learn and adapt quickly to change