LUT University

LAND OF THE CURIOUS

CT60A7650 – DATABASE SYSTEMS MANAGEMENT

BIG DATA MANAGEMENT

Lecture

Jiri Musto, D.Sc.

BIG DATA

- Big data is characterized by V's
  - Volume: The amount of data
  - Variety: Varied sources, types and formats
  - Velocity: How fast data is generated, collected and processed
  - Veracity: Inconsistencies and uncertainty in data
  - Value: Data should be transformed into something useful, valuable
- Comes in different formats
  - Structured
  - Semistructured
  - Unstructured

BIG DATA USE-CASES

- Future prediction
  - Predicting what the future will bring based on existing data
  - Behaviour patterns, trends, changes
- User / product / service analysis
  - Analysis of the current situation
  - Customer segementation, product/service improvement,
- Machine learning
  - Teach a machine/program to act according to existing data
  - Targeted advertisements, recommendations

NOSQL DBMS FOR BIG DATA

- Cassandra
  - Used by Netflix, Twitter, Facebook,
- HBase
  - Used by Spotify, Adobe, Yahoo!
- MongoDB
  - eBay, EA,
- Neo4j
  - Lyft, NBC News, U.S. Army

DATA WAREHOUSES AND DATA LAKES

- Massive amounts of data is gathered and stored
  - Warehouses organize data before it is stored, stored in a **database**
  - Lakes store data in natural format, stored in **data repository**
- Data warehouse
  - One large database gathering data from multiple sources
  - Management depends on the database and DBMS chosen
- Data lake
  - Can include databases and different files / folders
  - Management varies drastically depending on the sources
  - There are platforms for managing data lakes, such as Amazon S3

USING BIG DATA

- To use big data, it is highly recommended that you first identify what you need
  - Given the amount of data, collecting and processing everything will take time and space

Volume

- As data keep constantly changing, you may have to choose when to refresh database copies
  - Refresh as soon as changes happen
  - Refresh on intervals

Velocity

- Use storage formats that can be used by the end users and connected applications

Variety

- If you have multiple sources of data, make sure they are compatible

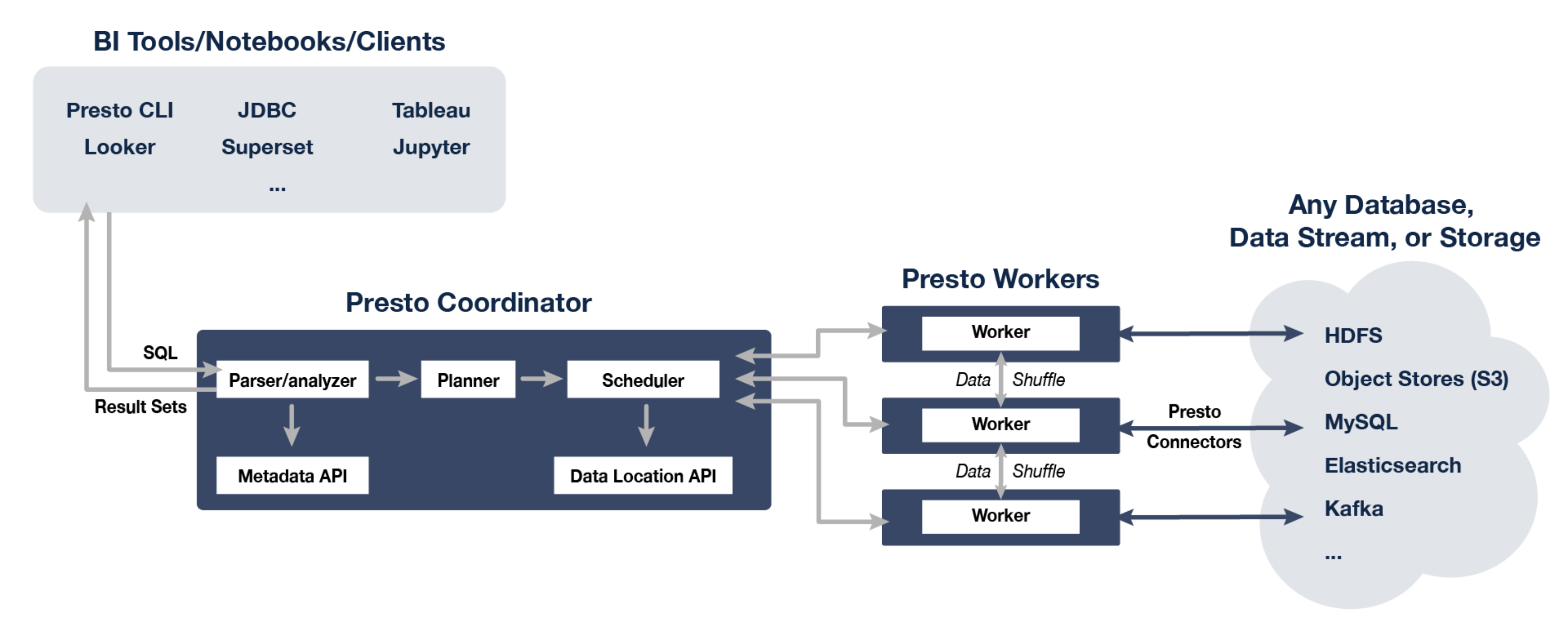
CHALLENGES WITH BIG DATA

- Large amount of data
- Problems with data quality
- Data integration
- Data preparation
- System scaling

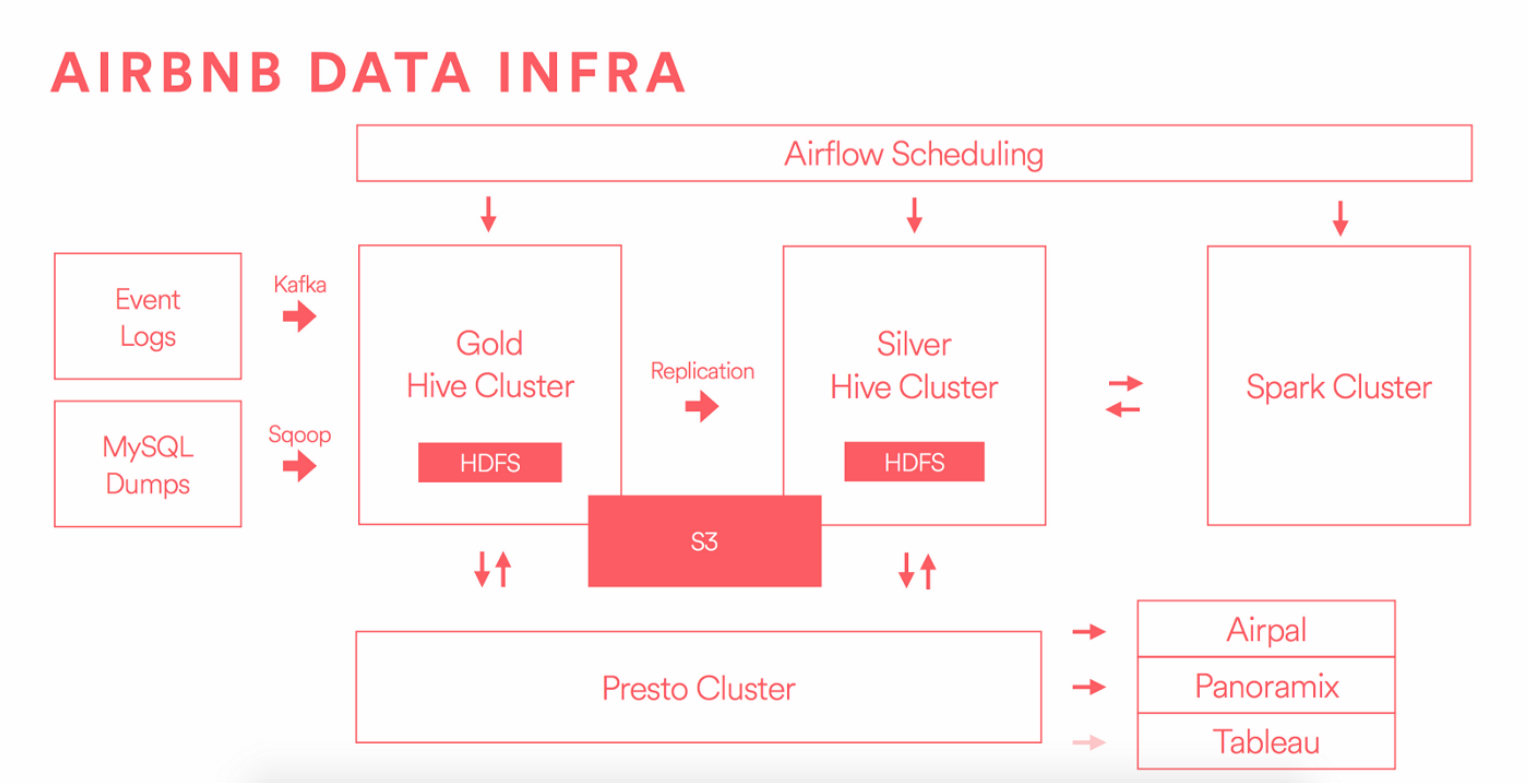
MANAGING BIG DATA

- Given the possible hardware limitations
  - Define how long the data should be stored
  - Define how much data is stored / collected
  - If you have a 5TB storage and are collecting 100GB every day, your storage will last 50 days
  - May be irrelevant in the future
- Data refreshing, how often?
  - If you retrieve data from sources
  - If you do data analysis based on big data
- Varied data formats
  - Define what is the "correct" format

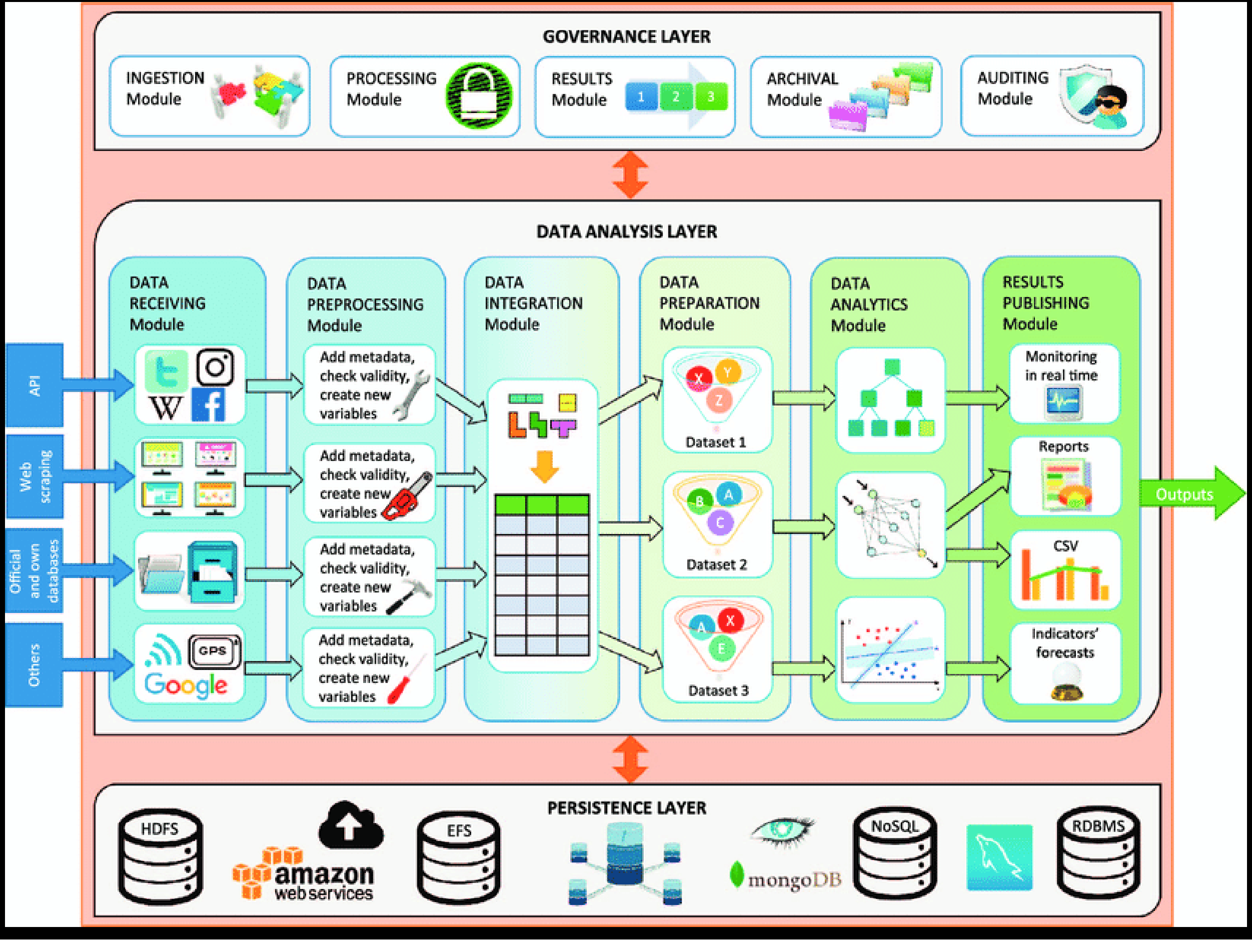
PRESTO: DISTRIBUTED SQL QUERY ENGINE FOR BIG DATA



AIRBNB DATA INFRA



BIG DATA ARCHITECTURE



GUIDELINES TO MANAGE BIG DATA

1. Create a detailed strategy from design to implementation and usage
2. Create a well-designed architecture
3. Focus on the business needs
4. Ensure data accessibility
5. Be flexible
6. Remember to handle access control

LUT
University