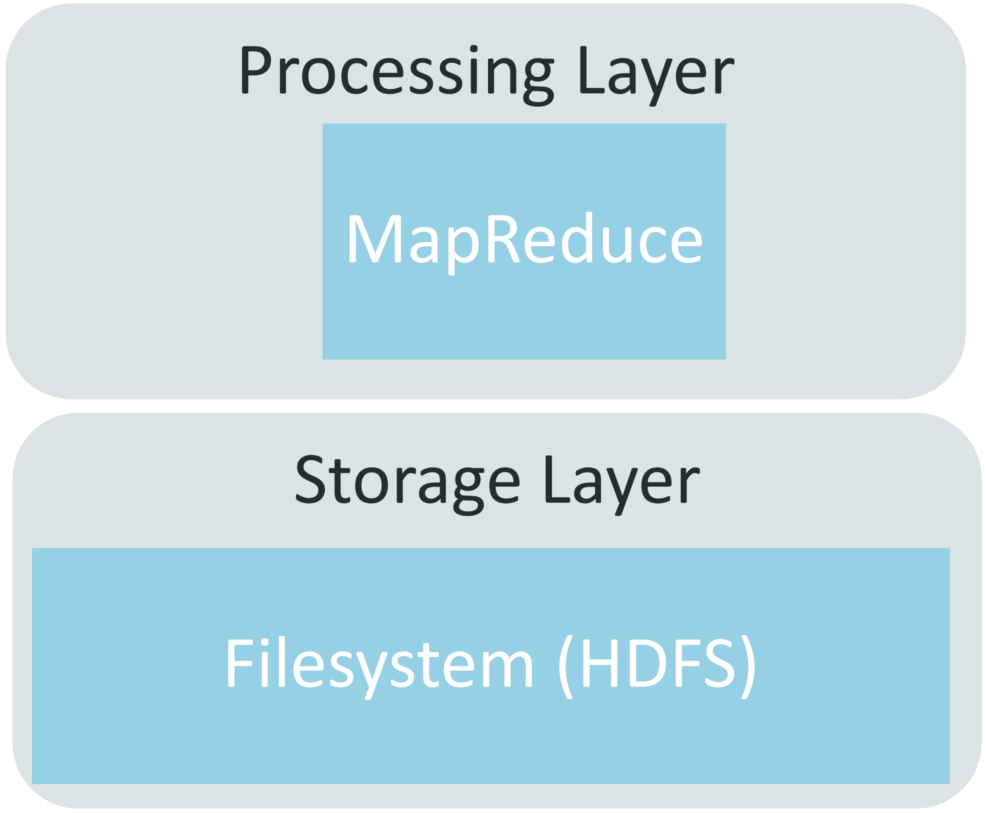
Hadoop basis and Hadoop Enviroment

Alexey Filanovskiy
Oracle Global DW and Big Data Leaders
Product Development server technologies

What is Hadoop?

What is Hadoop?

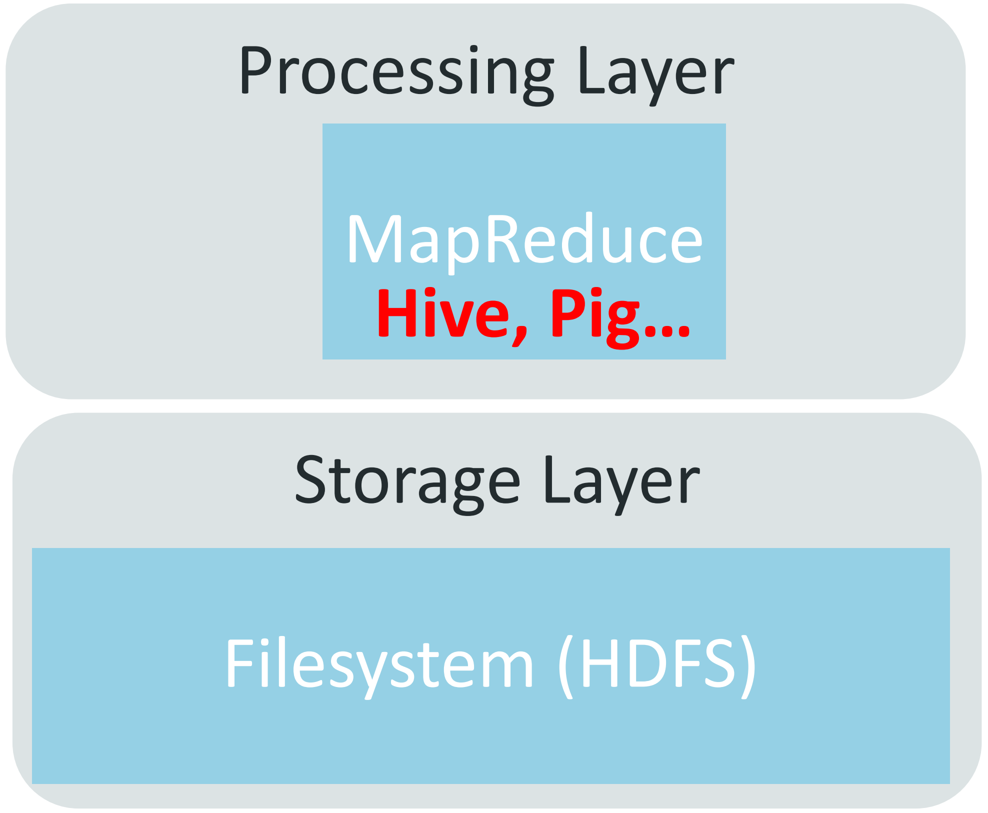


Historically (step 0 of evolution):
It was Distributed file system (logically you observe one directory, physically it lies on many servers) – Hadoop Distributed File System (HDFS)
+
Framework of parallel computations over this file system – MapReduce
Main goals of creation:
- Hight availability
- Scalability
Note: performance is optional requirement

Timeframe: Roughly 2006

What is Hadoop?

What is Hadoop? Hive

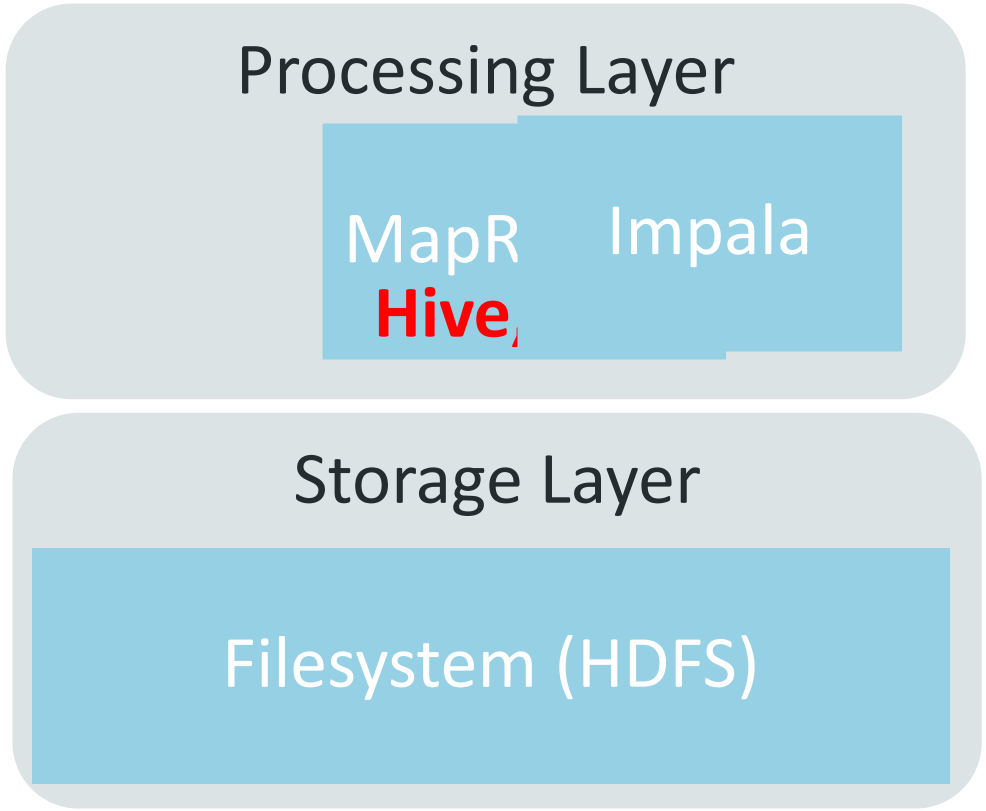


Step 1 of evolution:
Problem: MapReduce works fine, but for using it you have to write Java code. Instead this many users wanted hight level API (like SQL) for processing data that lies on HDFS

Solution: Special tools that convert high level API (like SQL) to Java MapReduce. Example of the tools: Hive, Pig

Timeframe: Roughly 2008

What is Hadoop? Impala

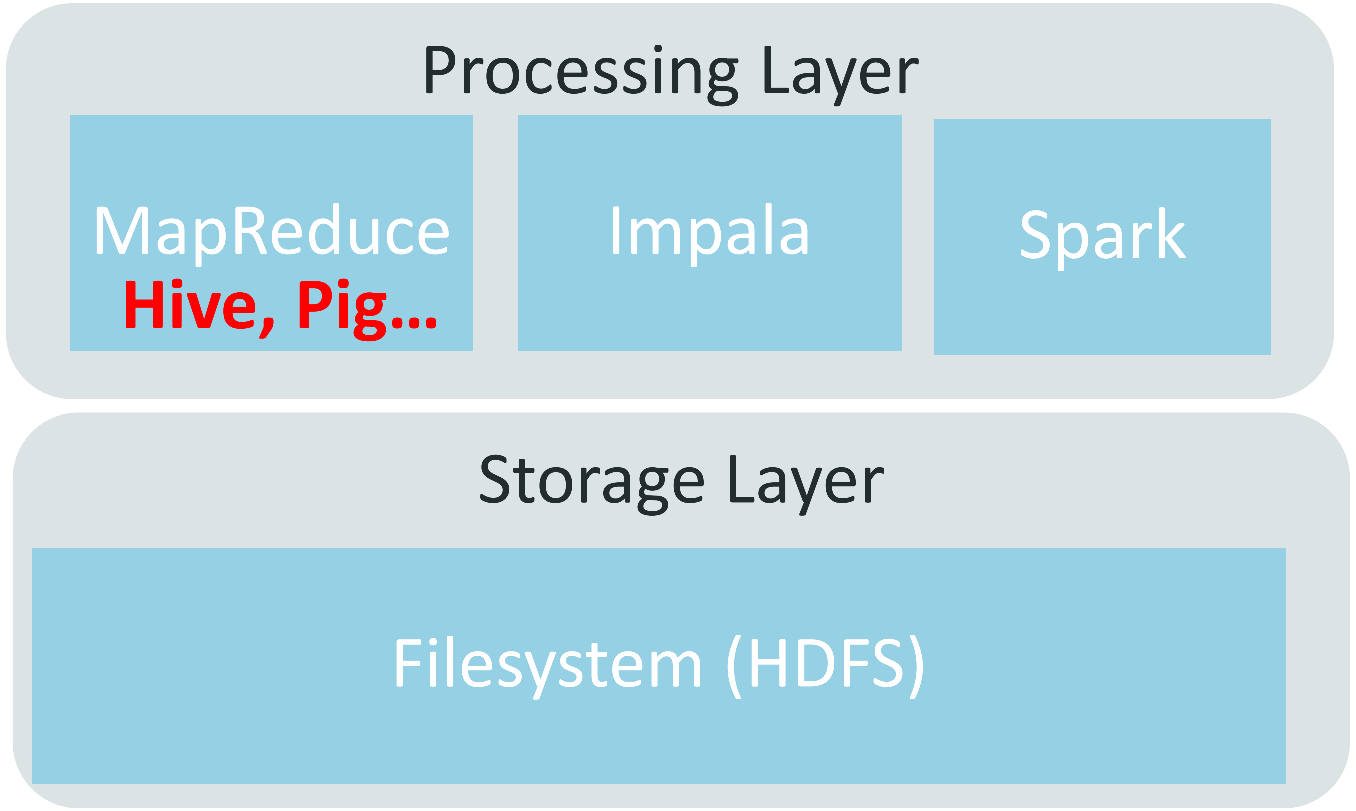


Step 2 of evolution:
Problem: Historically MapReduce was created for scalability and high availability.
Performance was not primary goal. Some of the users wanted to make sacrifice of scalability in favor of Performance. They also want to continue work with High level API (like SQL)

Solution: create special processing tool (NOT MAP REDUCE), that works over HDFS. Put in design of this tools performance considerations over reliability (like fast, but not reliable).

Timeframe: Roughly 2011

What is Hadoop? Spark



Step 3 of evolution:
Problem: MR was created in 2006. A lot of findings were done. It will be great to create something like new generation of MapReduce!

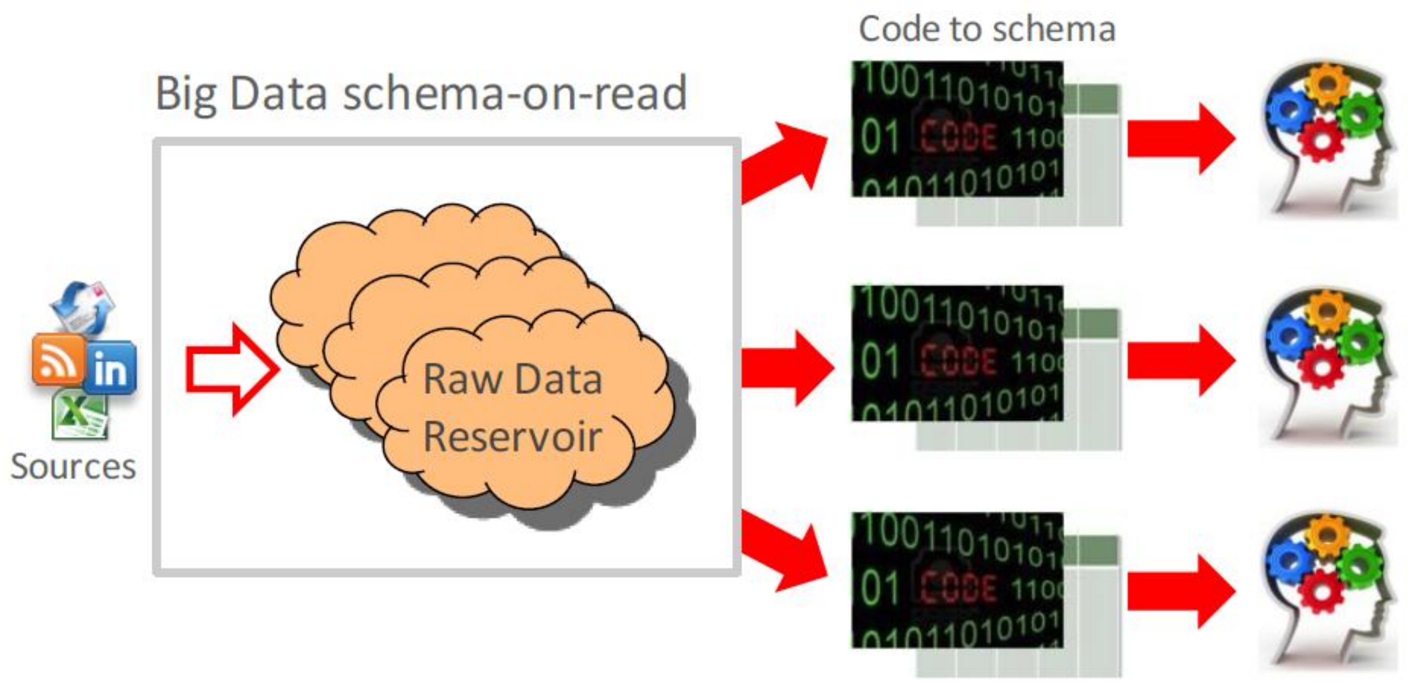
Solution: Spark has been developing since 2013 like new generation of MapReduce

Timeframe: Roughly 2013

And in all cases we are using benefits of Schema on Read approach (HDFS)...

Schema on read vs schema on write

Schema on read approach

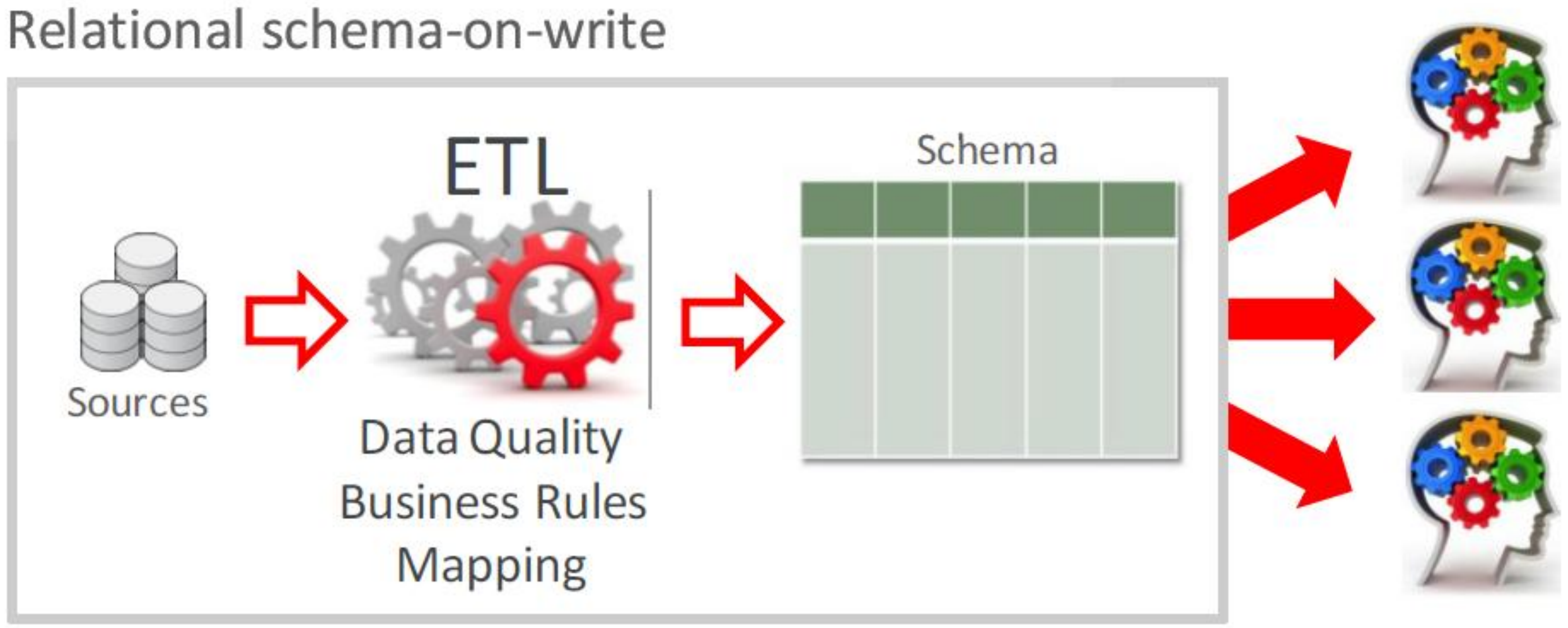


Key concept:
- Copy data "as is" like a source files on some file system (like coping files from flashcard to the laptop)
- Format will be defined later during the reading
- For different users the same data could be represented in different form

Advantages:
- High performance for data writing
- Highly flexible approach, it's no need to define format of the data for loading it

Disadvantages:
- Lower performance for end users. Every time they have to parse data

Schema on write approach



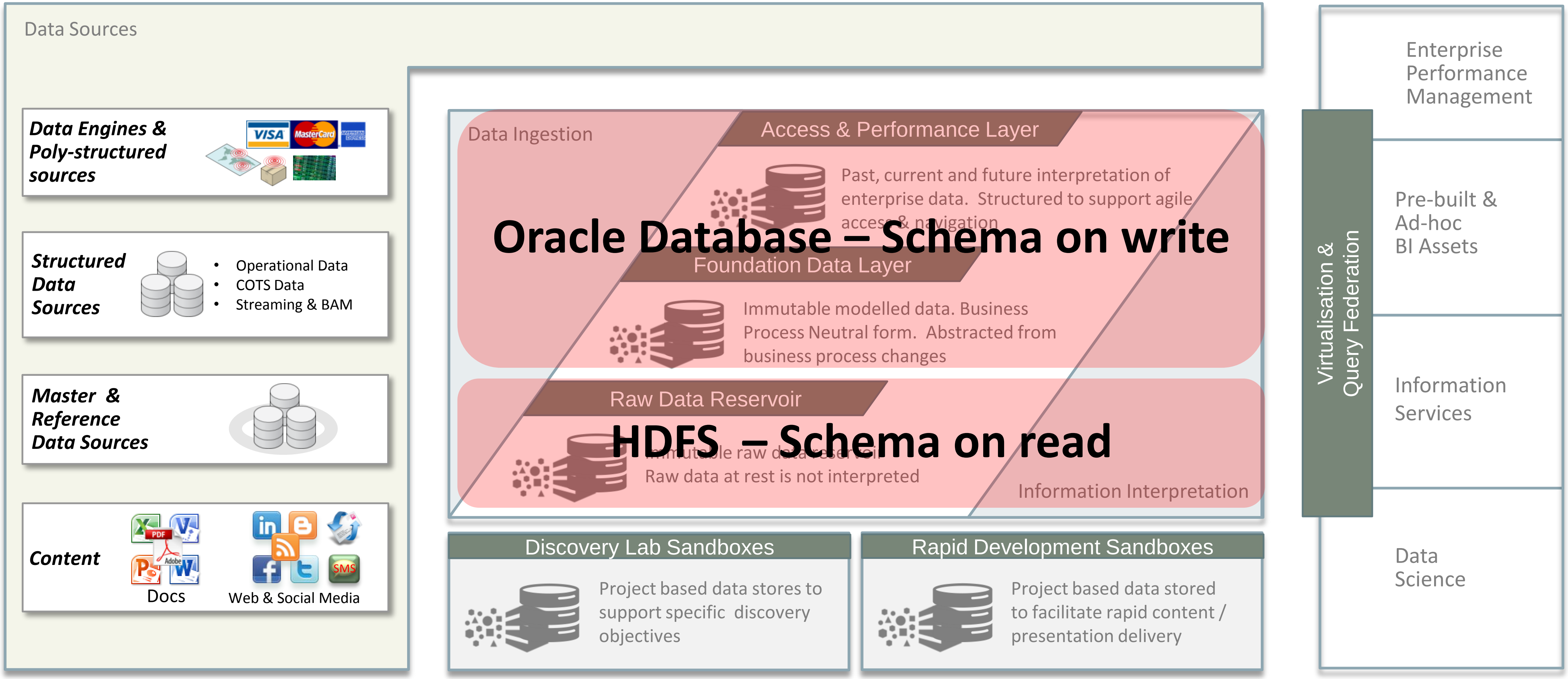
Key concept:
- Define columns
- Define datatypes
- Parse source data according the definition of your schema
- Load data into table
- All users read the same data

Advantages:
- High performance for end users
- Higher quality of the data comparable to schema on read approach

Disadvantages:
- Absence of flexibility
- Low performance for data loading

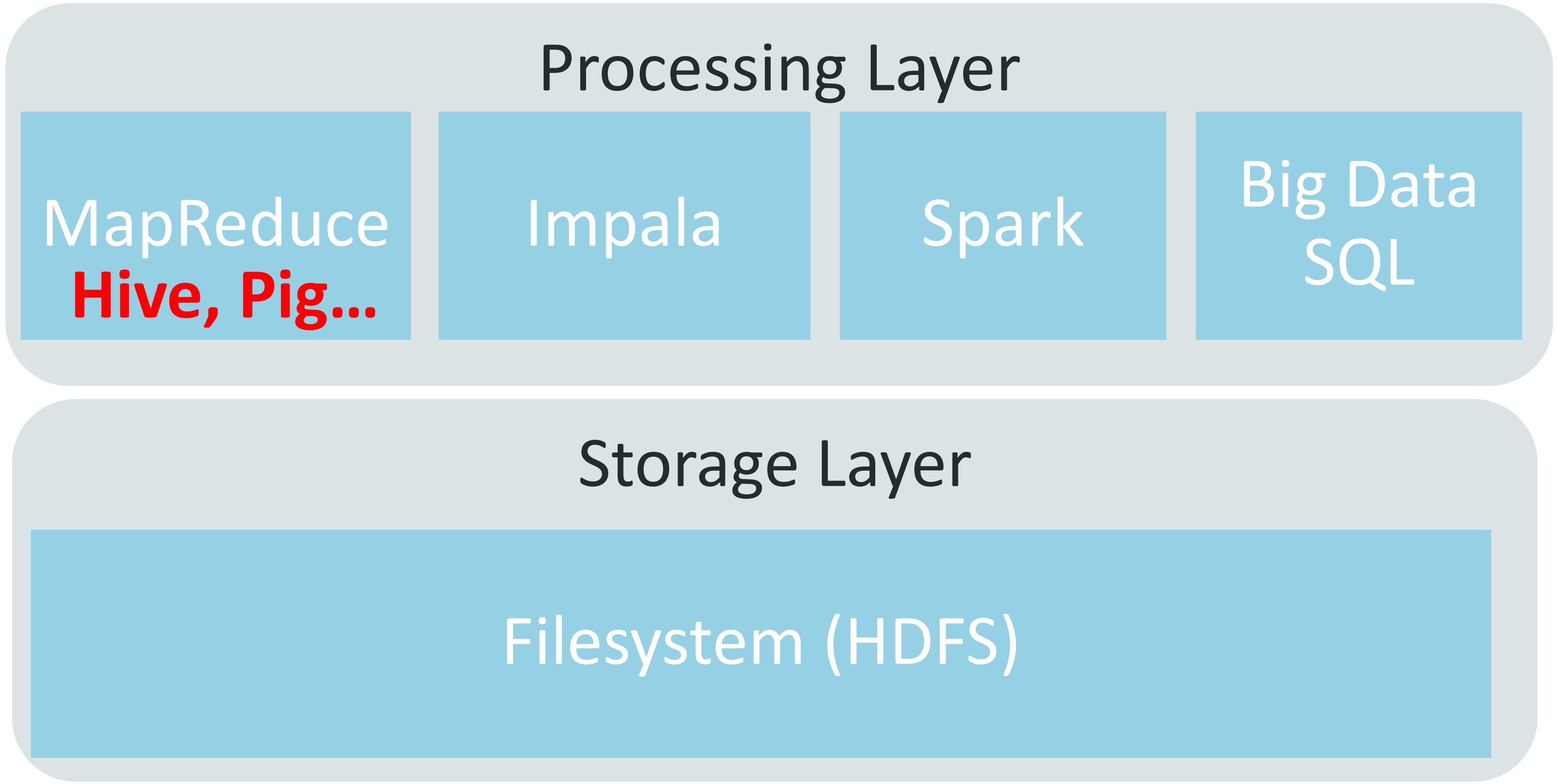
Who is good? Who is bad?

Reference architecture


Data Sources
Data Engines & Poly-structured sources
Structured Data Sources
Operational Data
COTS Data
Streaming & BAM
Master & Reference Data Sources
Content
Docs
Web & Social Media
Data Ingestion
Access & Performance Layer
Past, current and future interpretation of enterprise data. Structured to support agile access & navigation
Oracle Database – Schema on write
Foundation Data Layer
Immutable modelled data. Business Process Neutral form. Abstracted from business process changes
Raw Data Reservoir
HDFS – Schema on read
Immutable raw data reservoir
Raw data at rest is not interpreted
Information Interpretation
Discovery Lab Sandboxes
Project based data stores to support specific discovery objectives
Rapid Development Sandboxes
Project based data stored to facilitate rapid content / presentation delivery
Virtualisation & Query Federation
Enterprise Performance Management
Pre-built & Ad-hoc BI Assets
Information Services
Data Science

Big Data SQL

What is Hadoop? Big Data SQL



Step 4 of evolution:
Problem: Hadoop users understand value of schema on read approach, but they want tightly integrate it current Oracle relational database solution.

Solution: Big Data SQL

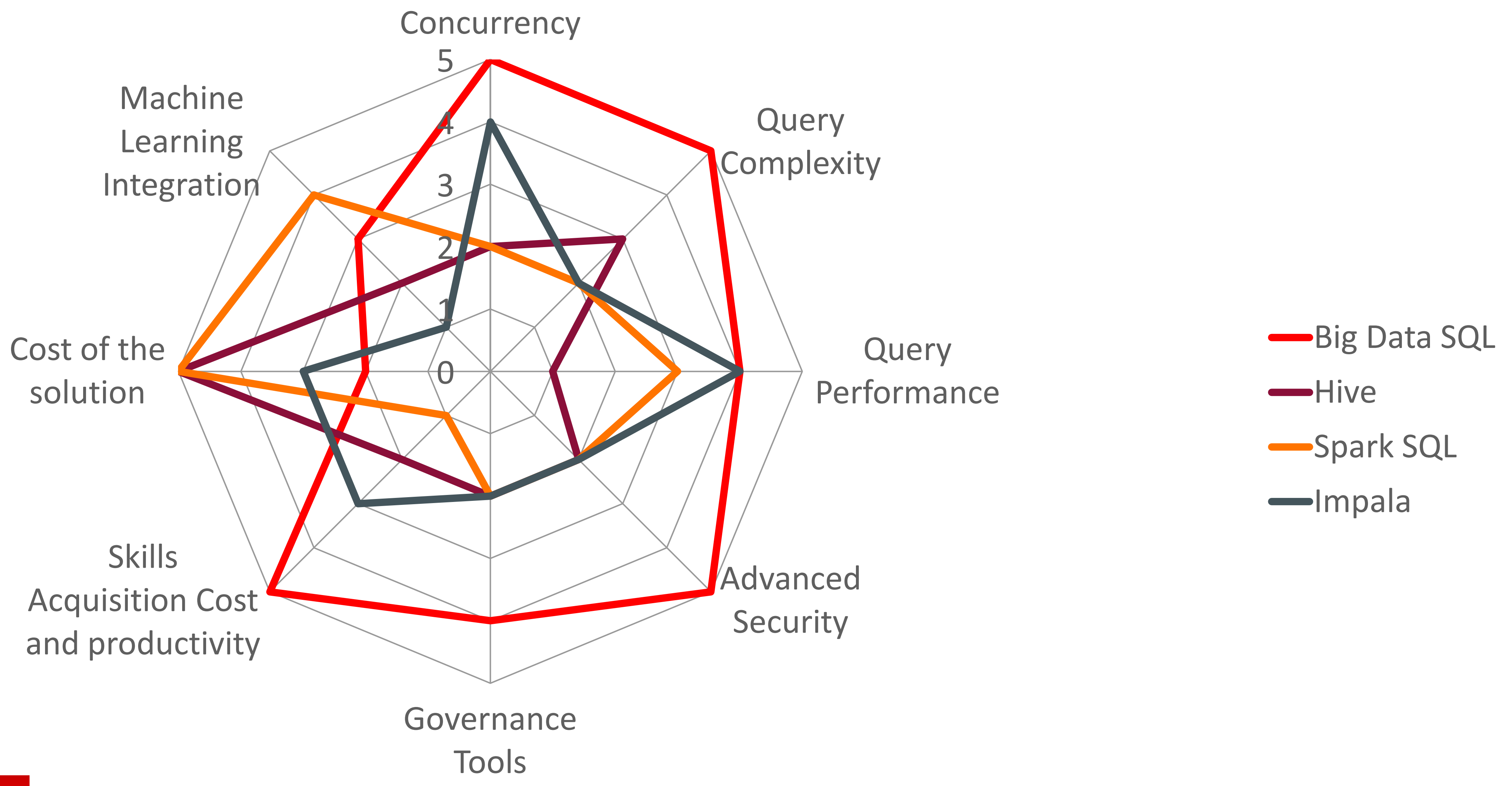
Timeframe: Roughly 2014

What is Hadoop? Conclusion.

Hadoop is:
- Sum of distributed file system (HDFS) and some processing framework
- As processing framework could be considered: MapReduce, Spark, Impala, Big Data SQL…
- In all cases it’s schema on read approach, that’s mean that we parse data during each read operation

Comparison

Comparison



Concurrency
Query Complexity
Query Performance
Advanced Security
Governance Tools
Skills Acquisition Cost and productivity
Cost of the solution
Machine Learning Integration
0
1
2
3
4
5
Big Data SQL
Hive
Spark SQL
Impala

Hardware and Software

Engineered to Work Together