Data Intensive Computing

Engin Arslan

earslan@unr.edu

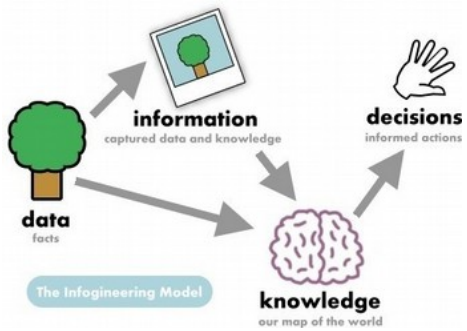
Jan 23, 2018

University of Nevada, Reno

Data are not much use without human intuition ...

Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom.

- Clifford Stoll



... analyzing data gives

Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway.

- Geoffrey Moore

Analyzing data is worth the cost ...

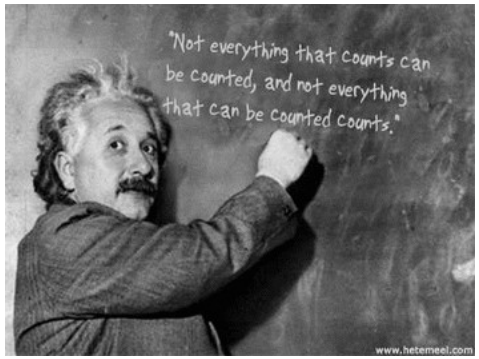
The price of light is less than the cost of darkness.

- Arthur C. Nielsen

..., but there are problems with relying on data too much.

Not everything that can be counted counts, and not everything that counts can be counted.

- Albert Einstein



Data is a treasure ..., except when it is not.

Getting information off the Internet is like taking a drink from a fire hose.

- Mitchell Kapor

However, any data is better than none.

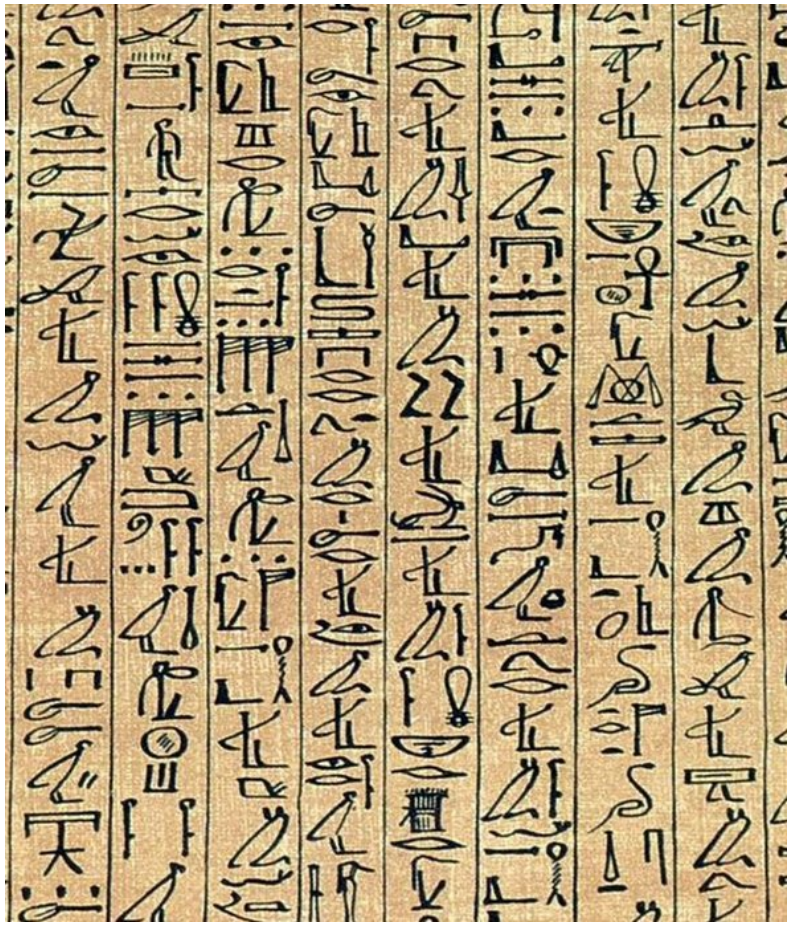
An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

- John Tukey

A Brief History of Data Management!

4000 B.C

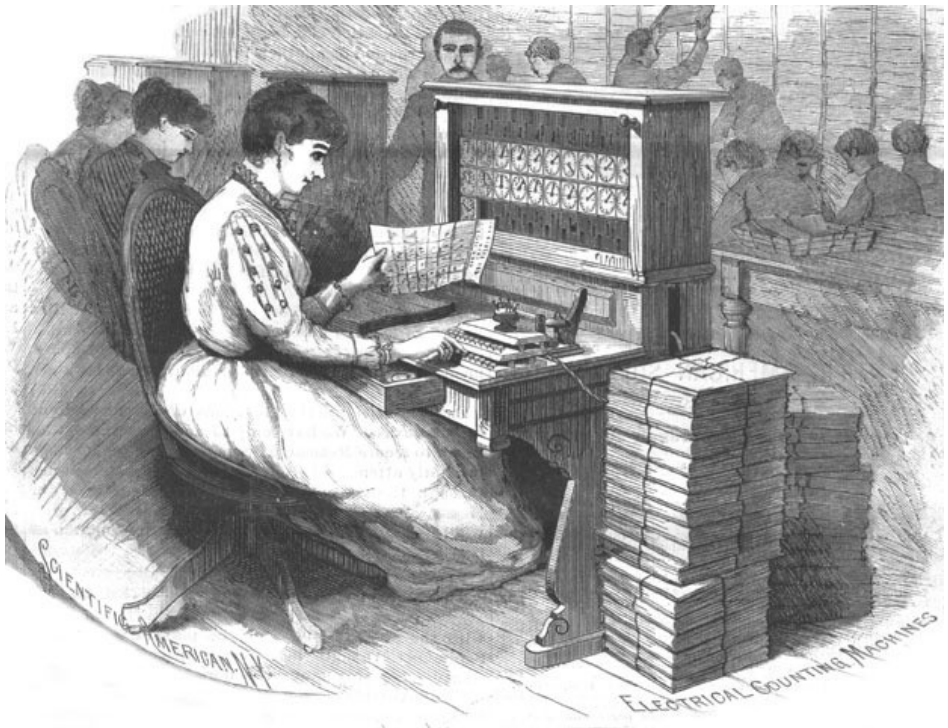
- Manual recording
- From tablets to papyrus, to parchment, and then to paper



- Gutenberg's printing press

1800's - 1940's

- Punched cards (no fault-tolerance)
- Binary data
- 1890: US census
- 1911: IBM appeared



1940's - 1970's

- Magnetic tapes
- Batch transaction processing
- File-oriented record processing model (e.g., COBOL)
- Hierarchical DBMS (one-to-many)
- Network DBMS (many-to-many)

1980's

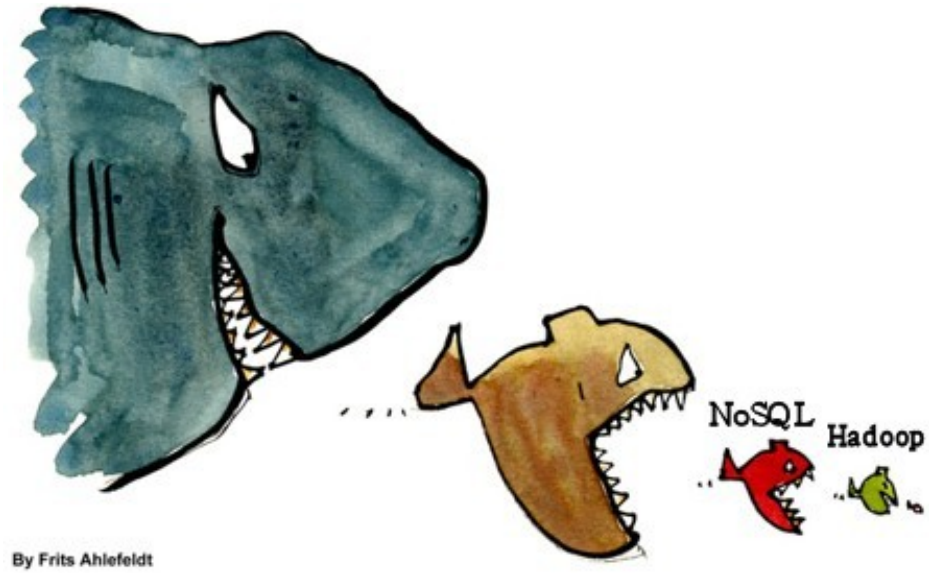
- Relational DBMS (tables) and SQL
- ACID
- Client-server computing
- Parallel processing

1990's - 2000's

- The Internet...

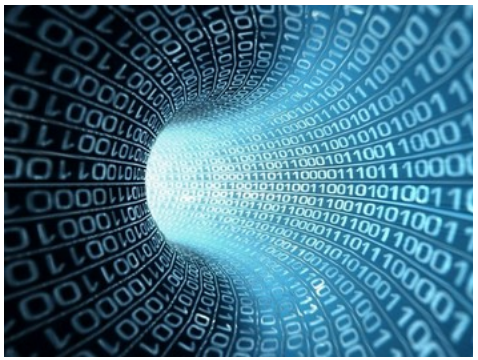
2010's

- NoSQL: BASE instead of ACID
- Big Data



Big Data

- In recent years we have witnessed a dramatic increase in available data.
- For example, the number of web pages indexed by Google, which were around one million in 1998, have exceeded one trillion in 2008, and its expansion is accelerated by appearance of the social networks.

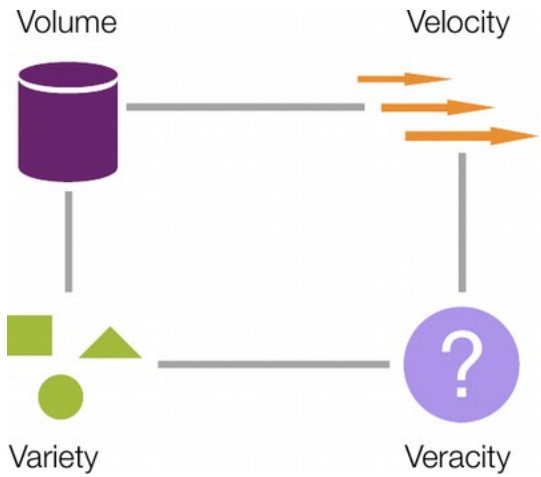


Big Data Definition

- Big Data refers to datasets and flows large enough that has outpaced our capability to store, process, analyze, and understand.

The Four Dimensions of Big Data

- Volume: data size
- Velocity: data generation rate
- Variety: data heterogeneity
- Veracity: uncertainty of accuracy and authenticity of data

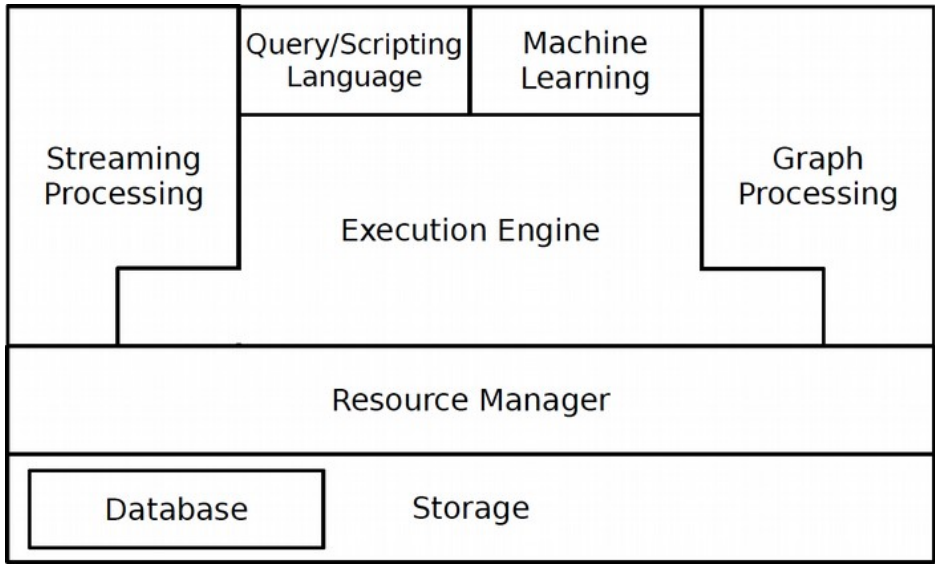


Big Data Market Driving Factors

- Mobile devices
- Internet of Things (IoT)
- Cloud computing
- Open source initiatives

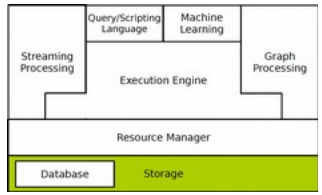
The Big Data Stack!

Big Data Analytics Stack



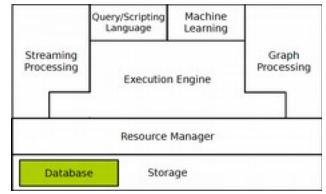
Big Data - Storage (Filesystem)

- Traditional filesystems are not well-designed for large-scale data processing systems.
- Efficiency has a higher priority than other features, e.g., directory service.
- Massive size of data tends to store it across multiple machines in a distributed way.
- HDFS, Amazon S3, ...



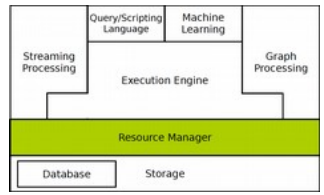
Big Data - Database

- Relational Databases Management Systems (RDMS)were not de- signed to be distributed.
- NoSQL databases relax one or more of theACID properties: BASE
- Different data models: key/value, column-family, graph, document.
- Dynamo, Scalaris, BigTable, Hbase, Cassandra, MongoDB, Voldemort, Riak, Neo4J, ...



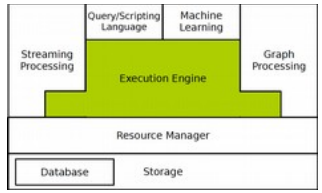
Big Data - Resource Management

- Different frameworks require different computing resources.
- Large organizations need the ability to share data and resources between multiple frameworks.
- Resource management share resources in a cluster between multiple frameworks while providing resource isolation.
- Mesos, YARN, Quincy, ...



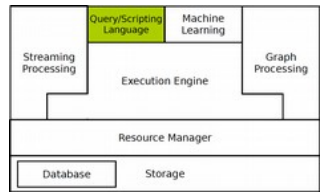
Big Data - Execution Engine

- Scalable and fault tolerance parallel data processing on clusters of unreliable machines.
- Data-parallel programming model for clusters of commodity machines.
- MapReduce, Spark, Stratosphere, Dryad, Hyracks, ...



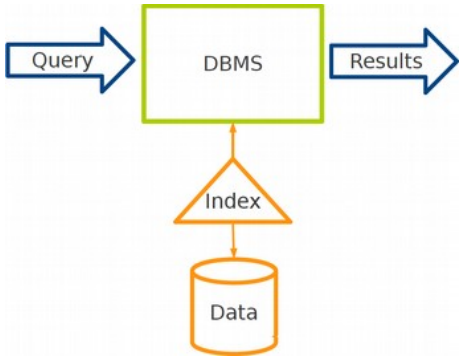
Big Data - Query/Scripting Language

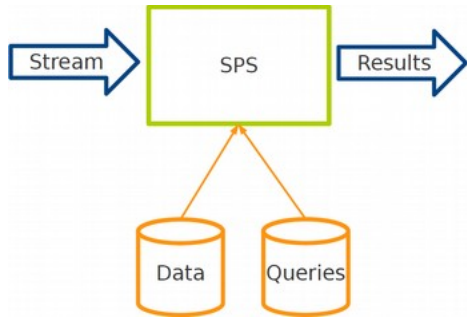
- Low-level programming of execution engines, e.g., MapReduce, is not easy for end users.
- Need high-level language to improve the query capabilities of exe- cution engines.
- It translates user-defined functions to low-level API of the execution engines.
- Pig, Hive, Shark, Meteor, DryadLINQ, SCOPE, ...



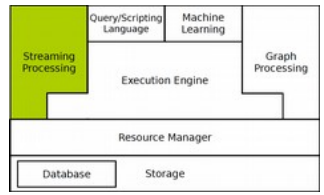
Big Data - Stream Processing

- Providing users with fresh and low latency results.
- Database Management Systems (DBMS) vs. Stream Processing Systems (SPS)



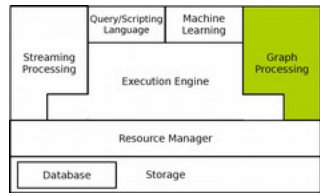


- Storm, S4, SEEP, D-Stream, Naiad, ...



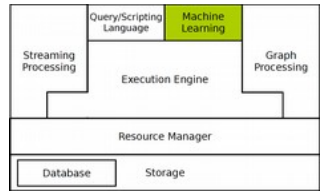
Big Data - Graph Processing

- Many problems are expressed using graphs: sparse computational dependencies, and multiple iterations to converge.
- Data-parallel frameworks, such as MapReduce, are not ideal for these problems: slow
- Graph processing frameworks are optimized for graph-based prob- lems.
- Pregel, Giraph, GraphX, GraphLab, PowerGraph, GraphChi, ...

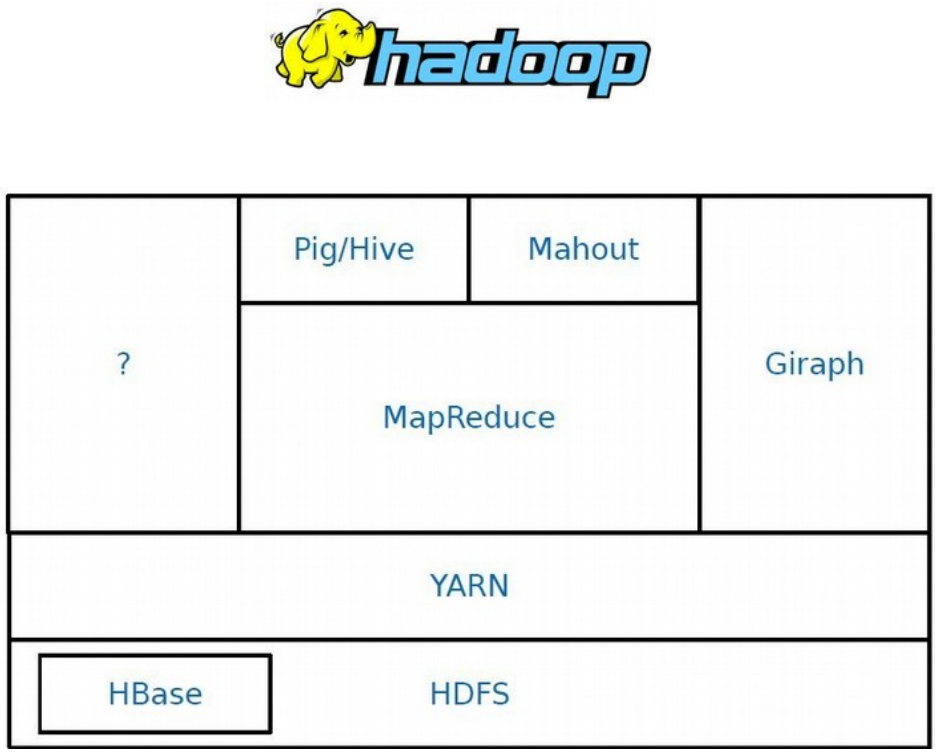


Big Data - Machine Learning

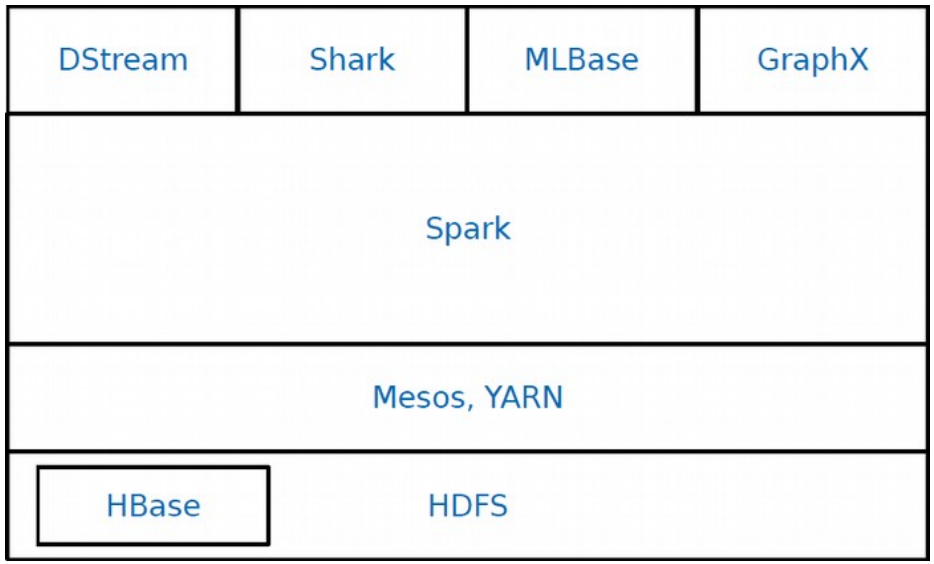
- Implementing and consuming machine learning techniques at scale are difficult tasks for developers and end users.
- There exist platforms that address it by providing scalable machine-learning and data mining libraries.
- Mahout, MLBase, SystemML, Ricardo, Presto, ...



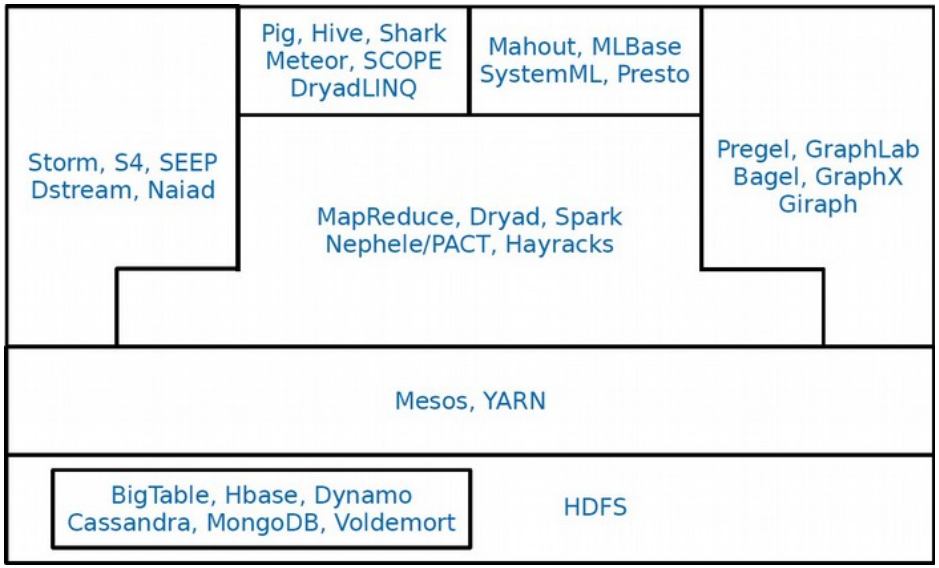
Hadoop Big Data Analytics Stack



Spark Big Data Analytics Stack



Summary


Pig, Hive, Shark
Meteor, SCOPE
DryadLINQ
Mahout, MLBase
SystemML, Presto
Storm, S4, SEEP
Dstream, Naiad
MapReduce, Dryad, Spark
Nephele/PACT, Hayracks
Pregel, GraphLab
Bagel, GraphX
Giraph
Mesos, YARN
BigTable, Hbase, Dynamo
Cassandra, MongoDB, Voldemort
HDFS

Questions?