cct | College Dublin
Computing • IT • Business

Big Data Storage and Processing

MSc in Data Analytics

CCT College Dublin

Revision BDSP

Week 13

Lecturer: Dr. Muhammad Iqbal*

Email: miqbal@cct.ie

Agenda

cct College Dublin
Computing • IT • Business

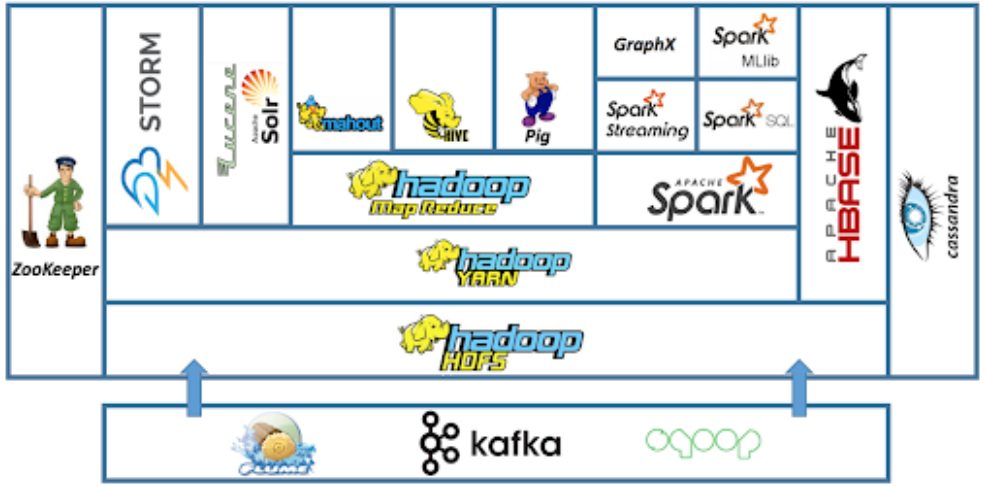
- Hadoop
- Map-Reduce Programming Model
- Apache HBase
- Apache Spark
- Apache Pig
- NOSQL Cassandra database
- YCSB & MongoDB
- Apache HIVE
- Graph Database
- Kafka Streaming

Hadoop Ecosystem

cct | College Dublin
Computing • IT • Business

- **Hadoop Ecosystem** is a platform which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions.

- **New Approach**
  - Google
  - Amazon
- Scalable Storage and Processing platforms based on commodity hardware
- Implemented as an open-source Apache Hadoop project



The Hadoop Ecosystem

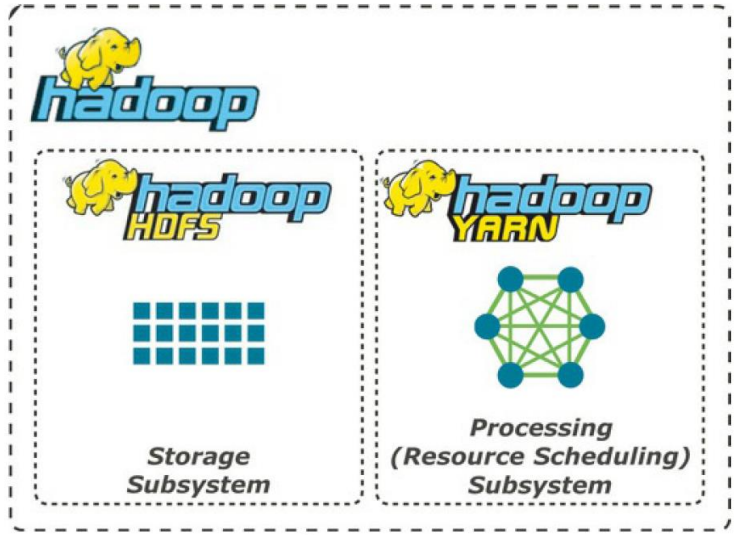
- **HDFS:** Hadoop Distributed File System
- **YARN:** Yet Another Resource Negotiator
- **MapReduce:** Programming based Data Processing
- **Spark:** In-Memory data processing

- **PIG, HIVE:** Query based processing of data services
- **HBase:** NoSQL Database
- **Mahout, Spark MLLib:** Machine Learning algorithm libraries
- **Zookeeper:** Managing cluster
- **Oozie:** Job Scheduling

Big Data and HDFS

cct College Dublin
Computing • IT • Business

- Scalable Storage and Processing platforms based on commodity hardware
  - **HDFS** – Hadoop Distributed File System
  - **MapReduce** is a programming model
- Storage of arbitrary data
  - Semi-structured
  - Unstructured
- Data interpreted at analysis stage
  - Allows for re-classification of data based on chosen algorithms as a part of analyses
- Optimised for large file storage
- Batch-oriented
- Supports streaming access to data
- Designed to answer the question: **"How to process big data with reasonable cost and time?"**



Hadoop core components.

Why the industry need Hadoop?

Big data brings with it two fundamental challenges

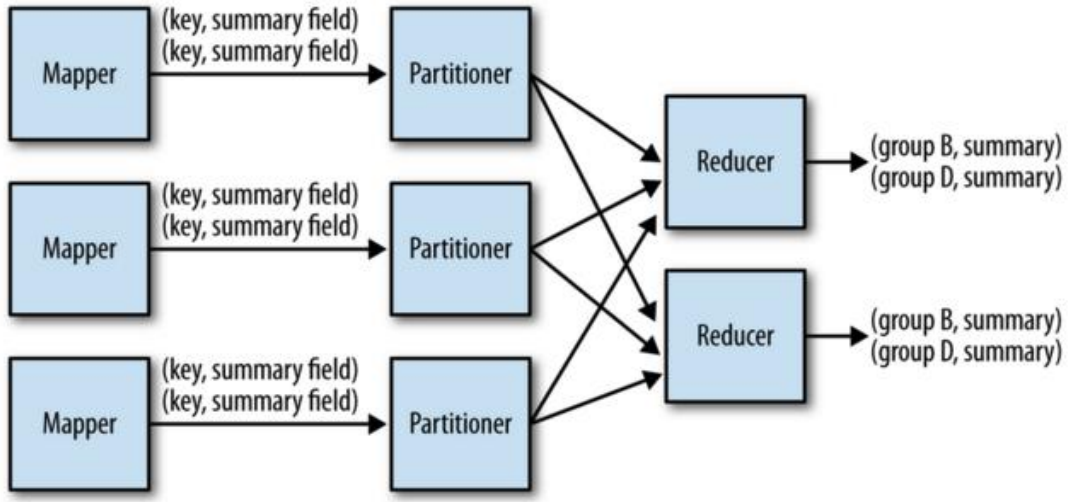
- How to store and work with voluminous data sizes?
- How to understand data and turn it into a competitive advantage?

Hadoop can do both tasks.

MapReduce Design Patterns

cct | College Dublin
Computing • IT • Business

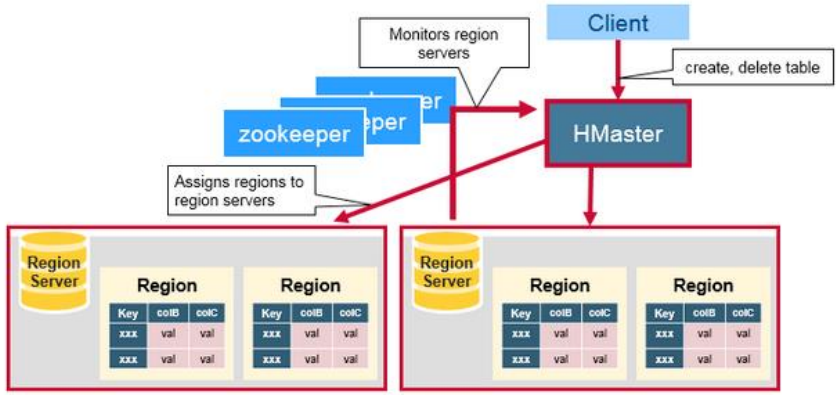
- **Categorization of Map-Reduce Patterns**
  - Summarization
  - Filtering
  - Join
  - Data Organization
  - Input /Output



Apache HBase

College Dublin
Computing • IT • Business

- HBase is a Java-based, open-source, non-relational distributed database that is based on Google's Bigtable.
- It is created as a component of the Apache Hadoop project and runs on top of HDFS or Alluxio, giving Hadoop capabilities similar to those of Bigtable.

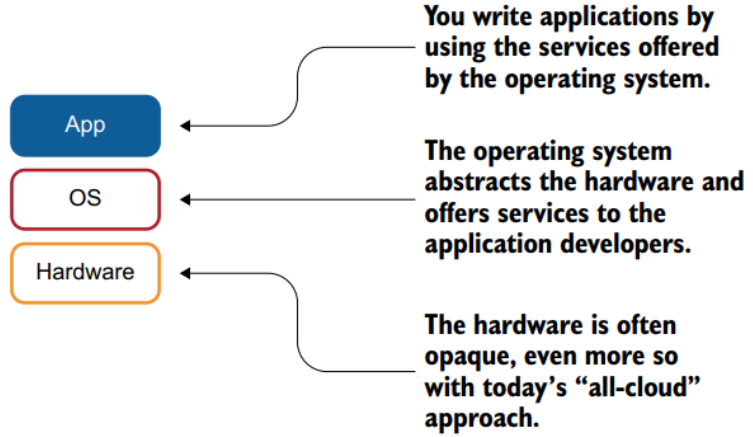


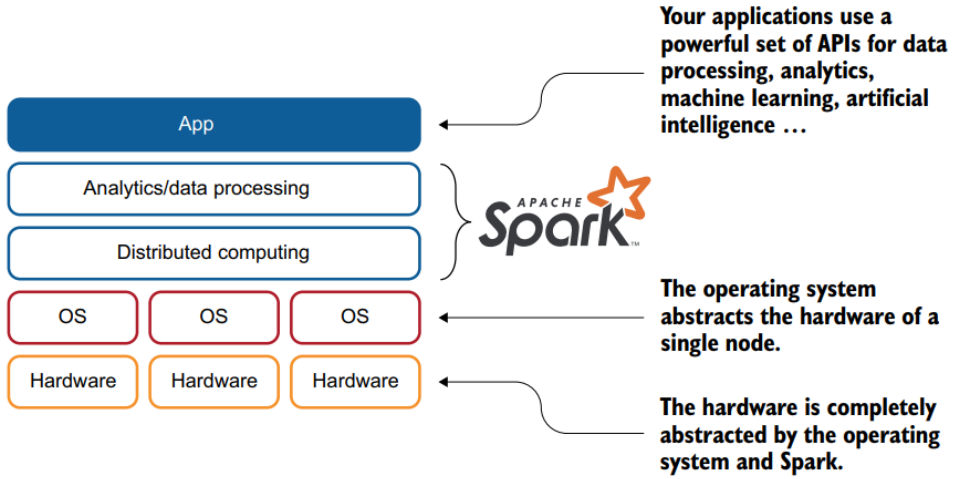
- The **master** is responsible for assigning regions to region servers and uses ***Apache Zookeeper***
  - Reliable, highly available, persistent, distributed coordination service
  - Necessary component for **HBase** operation

Apache Spark

cct College Dublin
Computing • IT • Business

- **Apache Spark** is a software stack for data scientists. When you build applications, you build them on top of an operating system as illustrated in the figure.
- The operating system provides services to make your application development easier; in other words, you are not building a filesystem or network driver for each application you develop.
- **Apache Spark** simplifies the development of analytics-oriented applications by offering services to applications as an operating system does.

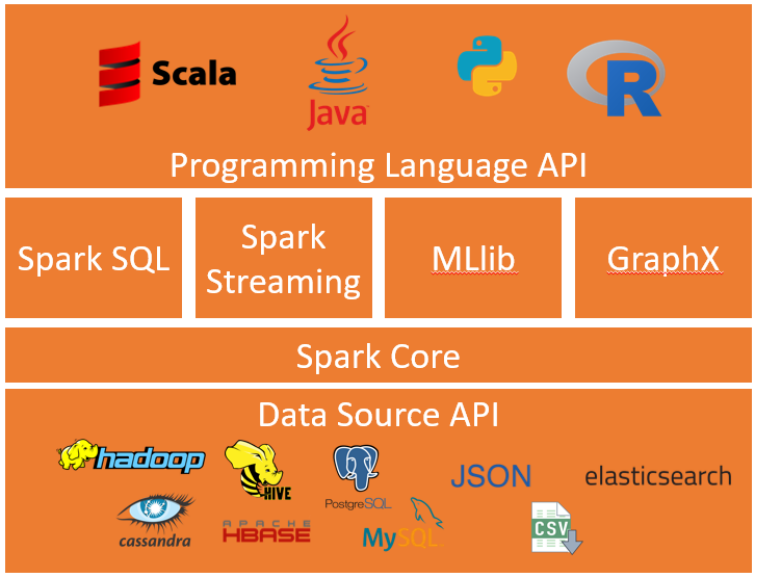


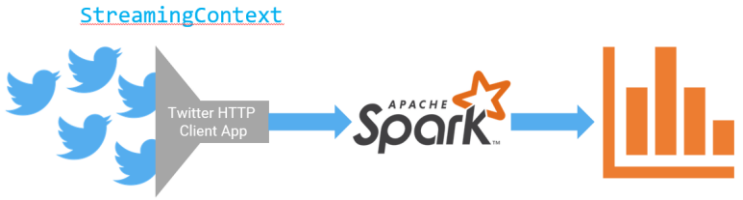


Apache Spark

cct College Dublin
Computing • IT • Business

- Data is generated continuously from one or many sources
- Sources can send the data simultaneously
- Data comes in small packages (kilobyte scale) in succession
- A lot of applications use continuously-updated data
- **Examples**
  - Sensors in vehicles, industrial equipment, and machinery send data to streaming for performance measurement.
  - Solar power company monitoring panel performance through streaming
  - Online gaming company collecting streaming data about player-game interactions
  - Spark streaming can use it to track most popular hashtags in 5 mins windows based on their counts in a Twitter stream, and by using the StreamingContext function.





Spark Programming Model

cct | College Dublin
Computing • IT • Business

- **Transformations**
- .filter(function)
- .distinct()
- .first()
- .union()
- .map(function)
- .flatMap(function)

- **Actions**
- .reduce()
- .count()
- .persist()
- .collect()
- .saveAsTextFile()
- .saveAsSequenceFile()

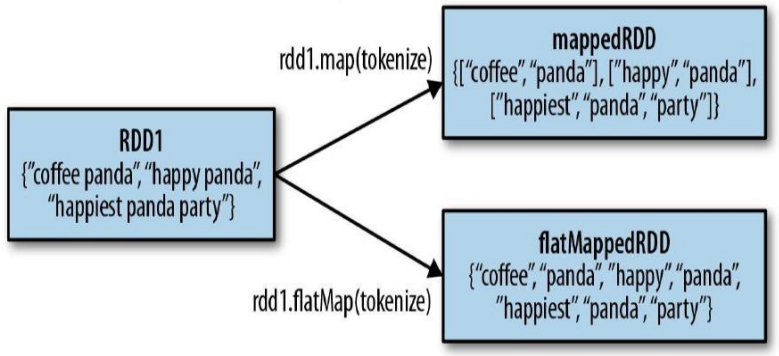


Figure 3-3. Difference between flatMap() and map() on an RDD

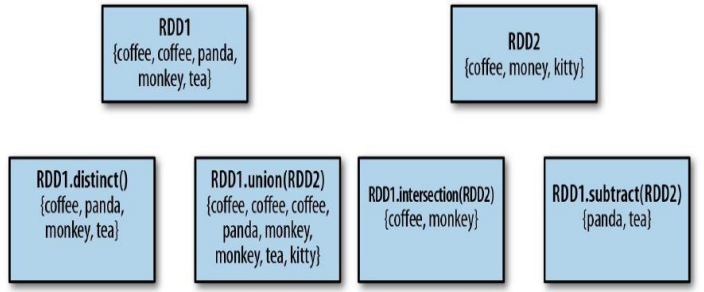
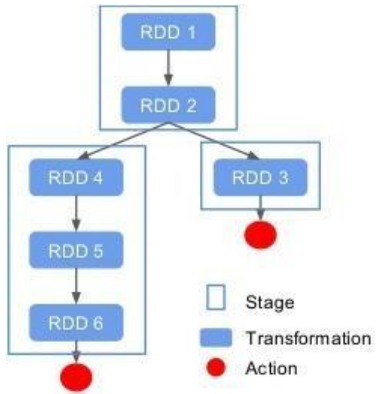


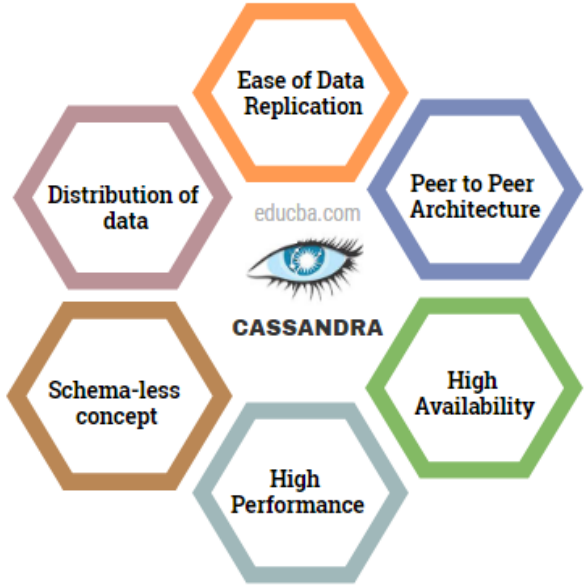
Figure 3-4. Some simple set operations



Apache Cassandra

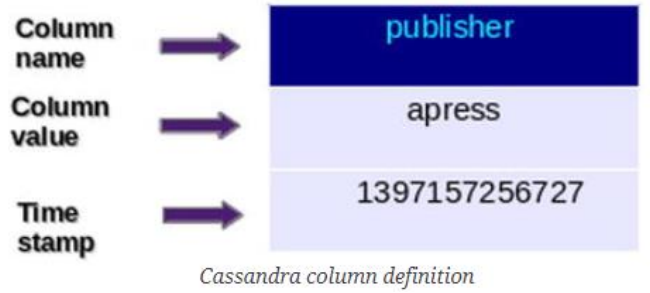
cct | College Dublin
Computing • IT • Business

- The **Apache Cassandra** data storage system differs greatly from a relational database management system in many ways.
- Apache Cassandra is an **open source, distributed, decentralized, elastically scalable, highly available, fault-tolerant, tuneable consistent, column-oriented database** that bases its distribution design on Amazon's Dynamo and its data model on Google's Bigtable.
- Created at Facebook and it is used at some of the most popular sites on the Web.
- In addition to performing blazingly fast writes, it can store hundreds of terabytes of data, and it is decentralized and symmetrical, so the failure point is eliminated. It provides schema-free data models and is highly available.
- Cassandra's interface allows it to be accessed from a variety of languages, including **C#, Scala, Python, and Ruby**.



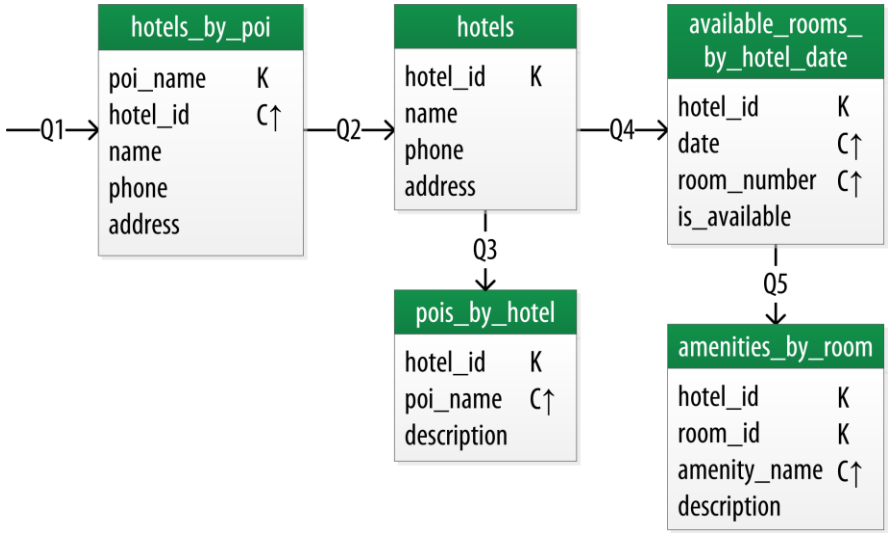
cassandra

Data Models Rules in Cassandra



Cassandra column definition

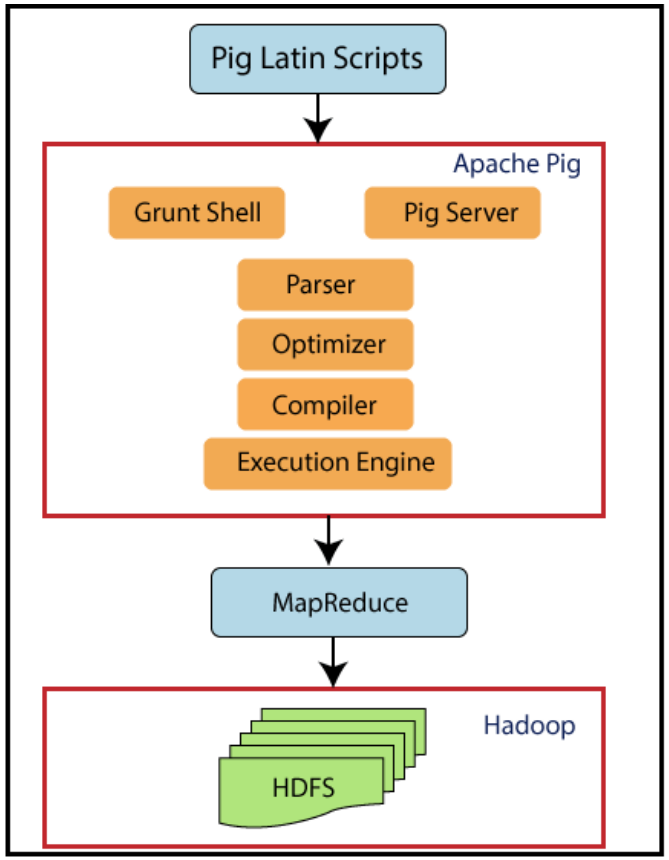
- **Cassandra** doesn't support ***JOINS, GROUP BY, OR*** clause, aggregation etc. You have to make sure the data is stored in a way that it can be accessed at any time.
- **Cassandra** is optimized for high **write** performances so you should maximize your writes for better read performance and data availability. There is a tradeoff between data write and data read. So optimize your data **read** performance by maximizing the number of data writes.
- Maximize data duplication because **Cassandra** is a distributed database and data duplication provides instant availability without a single point of failure.



- **Q1.** Find hotels near a given point of interest.
- **Q2.** Find information about a given hotel, such as its name and location.
- **Q3.** Find points of interest near a given hotel.
- **Q4.** Find an available room in a given date range.
- **Q5.** Find the rate and amenities for a room.

Apache Pig

- **Pig** is a client application
  - Requirement of a cluster is not essential
- Interprets **Pig** Latin scripts to **MapReduce** jobs
  - Parses Pig Latin scripts
  - Performs optimization
  - Performs compilation
  - Creates execution plan
- Submits **MapReduce** jobs to the **hdfs** cluster



Yahoo Cloud Serving Benchmark

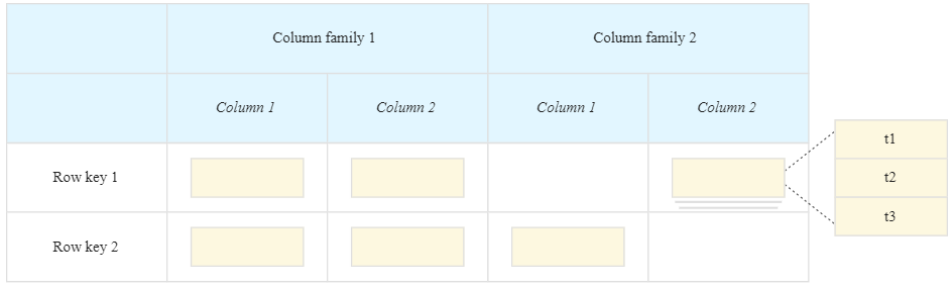
YCSB

College Dublin
Computing • IT • Business

- Most NoSQL databases make trade-offs like optimising for reads vs. writes, latency vs. durability, and synchronous vs. asynchronous replication, among others.
- The workloads they created for **YCSB** were created to "directly explore these trade-offs."
- These ***trade-offs*** are shown in Table below for a small number of databases.
- ***BigTable (Google)*** is designed for quick writes, long-term durability, and synchronous replication. It has a column-oriented data model.

System	Read/Write optimized	Latency/durability	Sync/async replication	Row/column
PNUTS	Read	Durability	Async	Row
BigTable	Write	Durability	Sync	Column
HBase	Write	Latency	Async	Column
Cassandra	Write	Tunable	Tunable	Column
Sharded MySQL	Read	Tunable	Async	Row

YCSB: Design decisions of various systems



https://cloud.google.com/bigtable/docs/overview

MongoDB

cct College Dublin
Computing • IT • Business

- **MongoDB** is a powerful, flexible, and scalable general-purpose database.
- **MongoDB** has the ability to scale out with features such as secondary indexes, range queries, sorting, aggregations, and geospatial indexes.
- **User-Friendliness**

- **MongoDB** is a **document-oriented** database, not a relational one. The primary reason for moving away from the relational model is to make scaling out easier, but there are some other advantages as well.
- The concept of a "row" is replaced by a more flexible model called "document" in a **document-oriented database, like MongoDB**.
- **MongoDB** allows complicated hierarchical relationships to be represented with just a single record by allowing embedded documents and arrays. This is how current object-oriented language developers think about their data.

Apache Hive

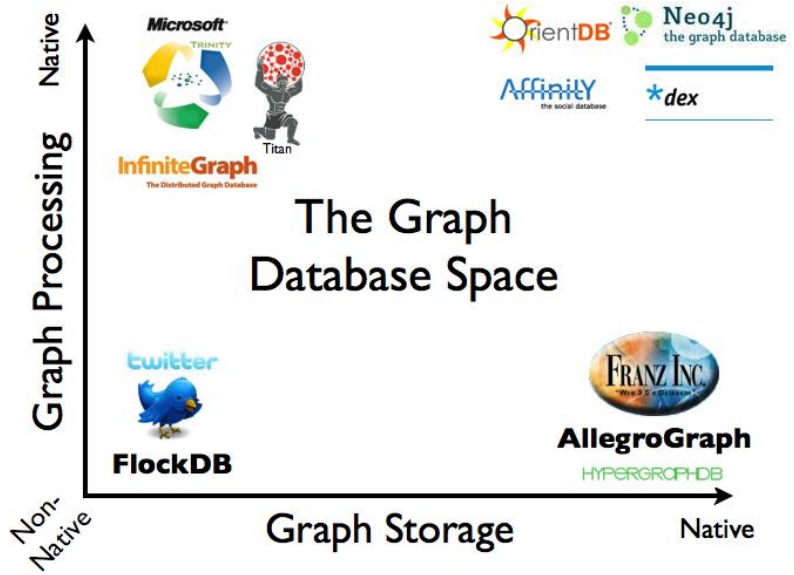
cct College Dublin
Computing • IT • Business

- Developed by Facebook and a top-level Apache project.
- A data warehousing infrastructure based on Hadoop.
- Immediately makes data on a cluster available to non-Java programmers via SQL like queries.
- Built on HiveQL (HQL), a SQL-like query language.
- Interprets HiveQL and generates MapReduce jobs that run on the cluster.
- Enables easy data summarization, ad-hoc reporting and querying, and analysis of large volumes of data.

Graph Databases

cct | College Dublin
Computing • IT • Business

- Graph data model treats relationships as first-class citizens.
- As compared to other DBMSs, we have to infer connections between entities using things like foreign keys or out-of-band processing such as map-reduce.
- By assembling the simple abstractions of nodes and relationships into connected structures, graph databases enable us to build sophisticated models that map closely to our problem domain.
- The resulting models are simpler and at the same time more expressive than those produced using traditional relational databases and the other NOSQL (Not Only SQL) stores.



- Figure shows a pictorial overview of some of the graph databases on the market today, based on their storage and processing models.

Kafka Streaming

cct College Dublin
Computing • IT • Business

- A client library called ***Kafka Streams*** is used to create ***apps*** and ***microservices*** whose input and output data are stored in Kafka clusters.
- It combines the advantages of Kafka's server-side cluster technology with the ease of creating and deploying regular Java and Scala apps on the client side.
- **Kafka Streams Use Cases**
- **The New York Times** uses Apache Kafka and the Kafka Streams to store and distribute, in real-time, published content to the various applications and systems that make it available to the readers.
- As of 2017, **trivago** offer access to approximately 1.8 million hotels and other accommodations in over 190 countries. **trivago** uses Kafka, Kafka Connect, and Kafka Streams to enable our developers to access data freely in the company. Kafka Streams powers parts of our analytics pipeline and delivers endless options to explore and operate on the data sources we have at hand.
- **Rabobank** is one of the 3rd largest banks in the Netherlands. Its digital nervous system, the Business Event Bus, is powered by Apache Kafka. It is used by an increasing amount of financial processes and services, one of which is ***Rabo Alerts***. This service alerts customers in real-time upon financial events and is built using Kafka Streams.

Resources/ References

- G. Somasundaram, A Shrivastava (Editors), 2009, Information Storage and Management: Storing, Managing, and Protecting Digital Information, Wiley Publishing
- Tate, J. et al, (2012). Introduction to Storage Area Networks and System Networking. IBM.
- Some images are used from Google search repository (https://www.google.ie/search) to enhance the level of learning.
- https://kafka.apache.org/documentation/streams/