# 01
# Introduction To Big Data

Demi Ben-Ari

# What are Big Data Systems?

- Applications involving the “3 Vs”
  - **Volume** - Gigabytes, Terabytes, Petabytes...
  - **Velocity** - Sensor Data, Logs, Data Streams, Financial Transactions, Geo Locations..
  - **Variety** - Different data format ingestion
- Some define it the “7 Vs”
  - **Variability** (constantly changing)
  - **Veracity** (accuracy)
  - **Visualization**
  - **Value**
- Characteristics
  - Multi-region availability
  - Very fast and reliable response
  - No single point of failure

# **What is Big Data** (IMHO)**?**

- Systems involving the **“3 Vs”**:

  What are the right questions we want to ask?

  - **Volume** - **How much?**
  - **Velocity** - **How fast?**
  - **Variety** - **What kind?** (Difference)

# Why Not Relational Data

- Relational Model Provides
  - Normalized table schema
  - Cross table joins
  - ACID compliance (**Atomicity, Consistency, Isolation, Durability**)
- But at very high cost
  - Big Data table joins - bilions of rows or more - require massive overhead
  - Sharding tables across systems is complex and fragile
- Modern applications have different priorities
  - Needs for speed and availability come over consistency
  - Commodity servers racks trump massive high-end systems
  - Real world need for transactional guarantees is limited

# What strategies help manage Big Data?

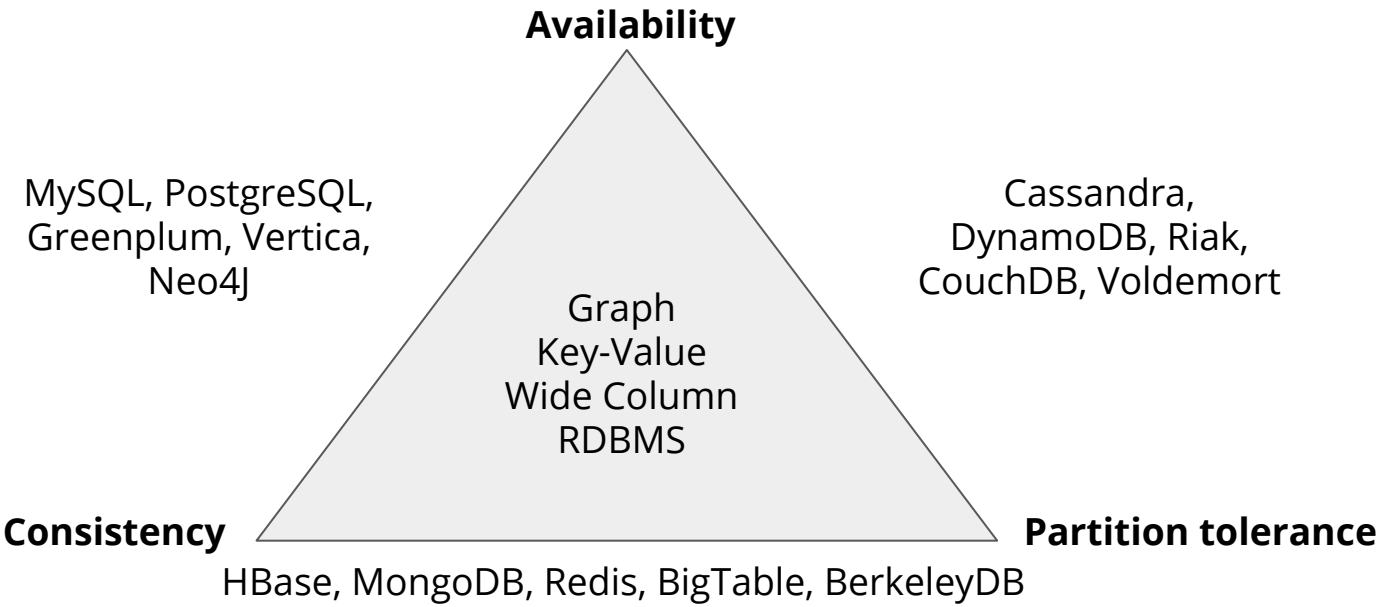
- Distribute data across nodes
  - Replication
- Relax consistency requirements
- Relax schema requirements
- Optimize data to suit actual needs

# What is the NoSQL landscape?

- 4 broad classes of non-relational databases (http://db-engines.com/en/ranking)
  - **Graph:** data elements each relate to N others in graph / network
  - **Key-Value:** keys map to arbitrary values of any data type
  - **Document:** document sets (JSON) queryable in whole or part
  - **Wide column Store** (Column Family): keys mapped to sets of n-numbers of typed columns
- Three key factors to help understand the subject
  - **Consistency:** do you get identical results, regardless which node is queried?
  - **Availability:** can the cluster respond to very high read and write volumes?
  - **Partition tolerance:** is a cluster still available when part of it is down?

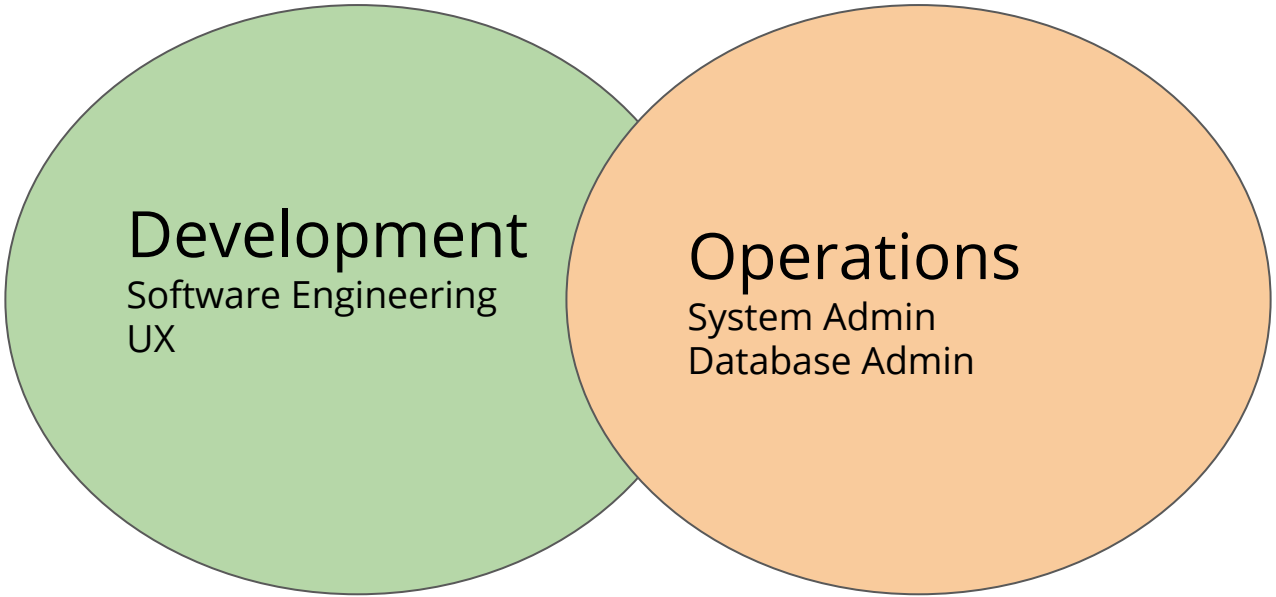
# What is the CAP theorem?

- In distributed systems, consistency, availability and partition tolerance exist in a manually dependant relationship, **Pick any two**.



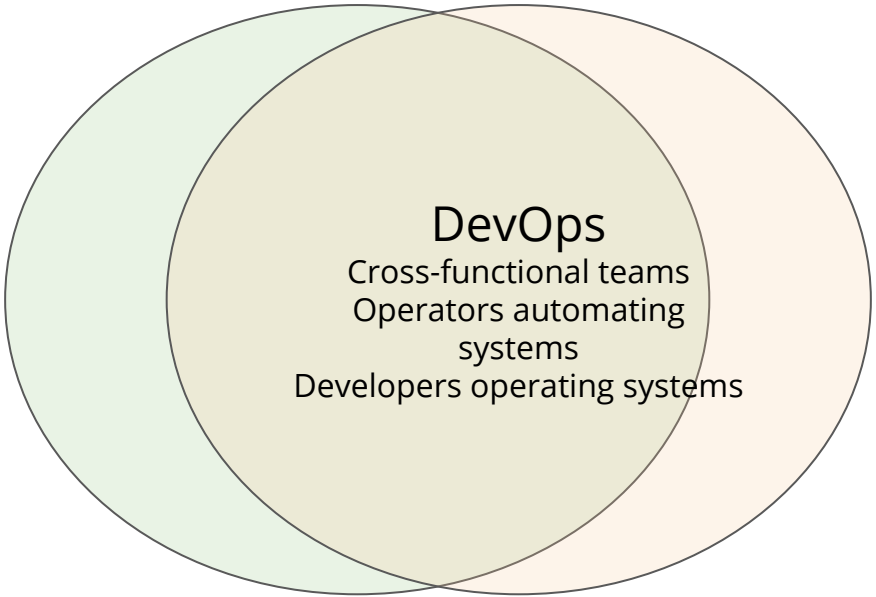

Panorays

# What does DevOps really mean?




Panorays

# What does DevOps really mean?



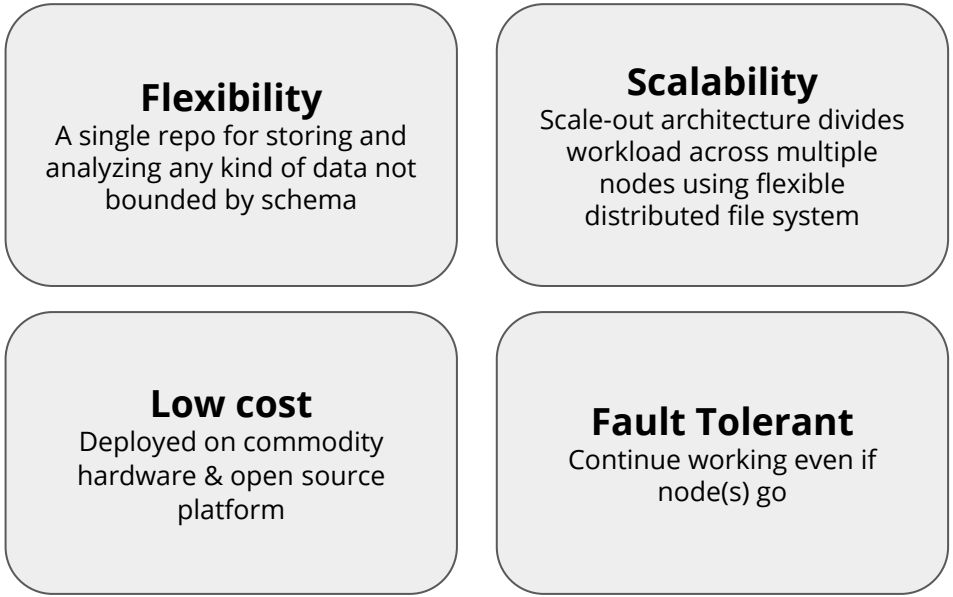

Panorays

# Characteristics of Hadoop

- A system to process very large amounts of unstructured and complex data with wanted speed
- A system to run on a large amount of machines that don’t share any memory or disk
- A system to run on a cluster of machines which can put together in relatively lower cost and easier maintenance

# Hadoop Principal

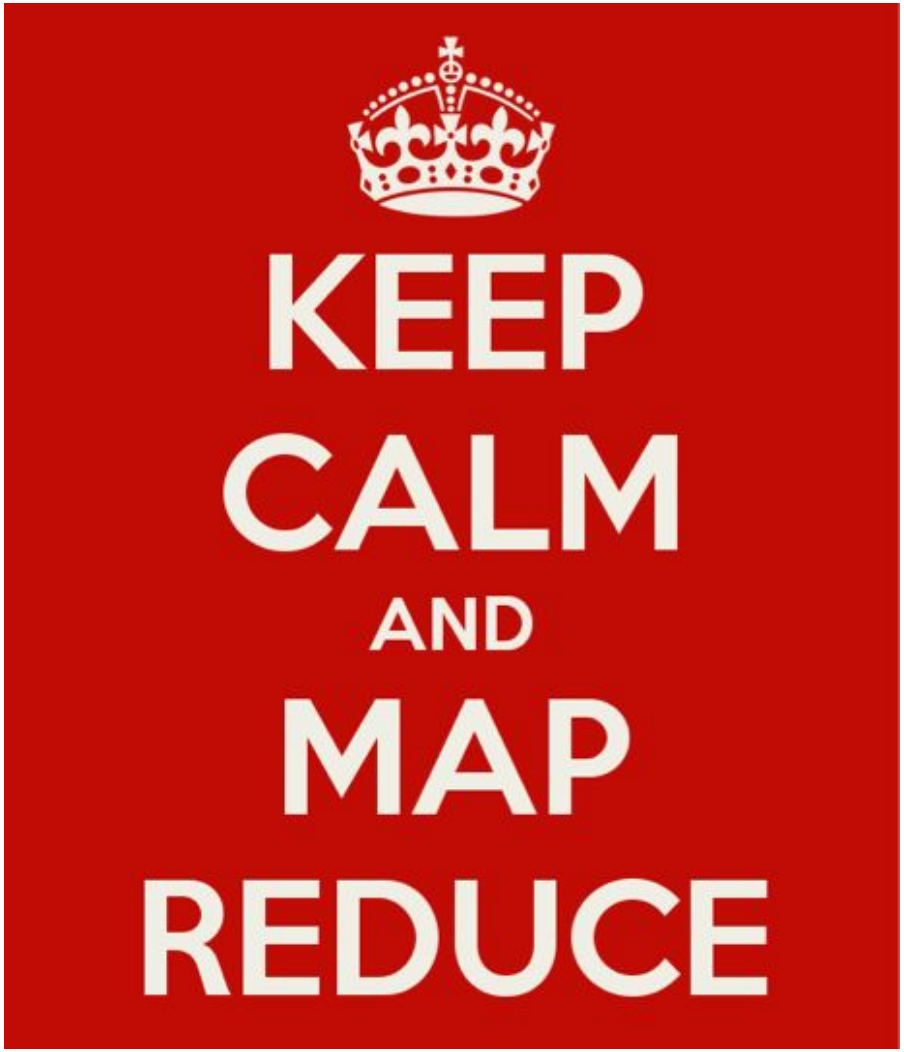
- “A system to move the computation, where the data is”
- Key Concepts of Hadoop




Panorays

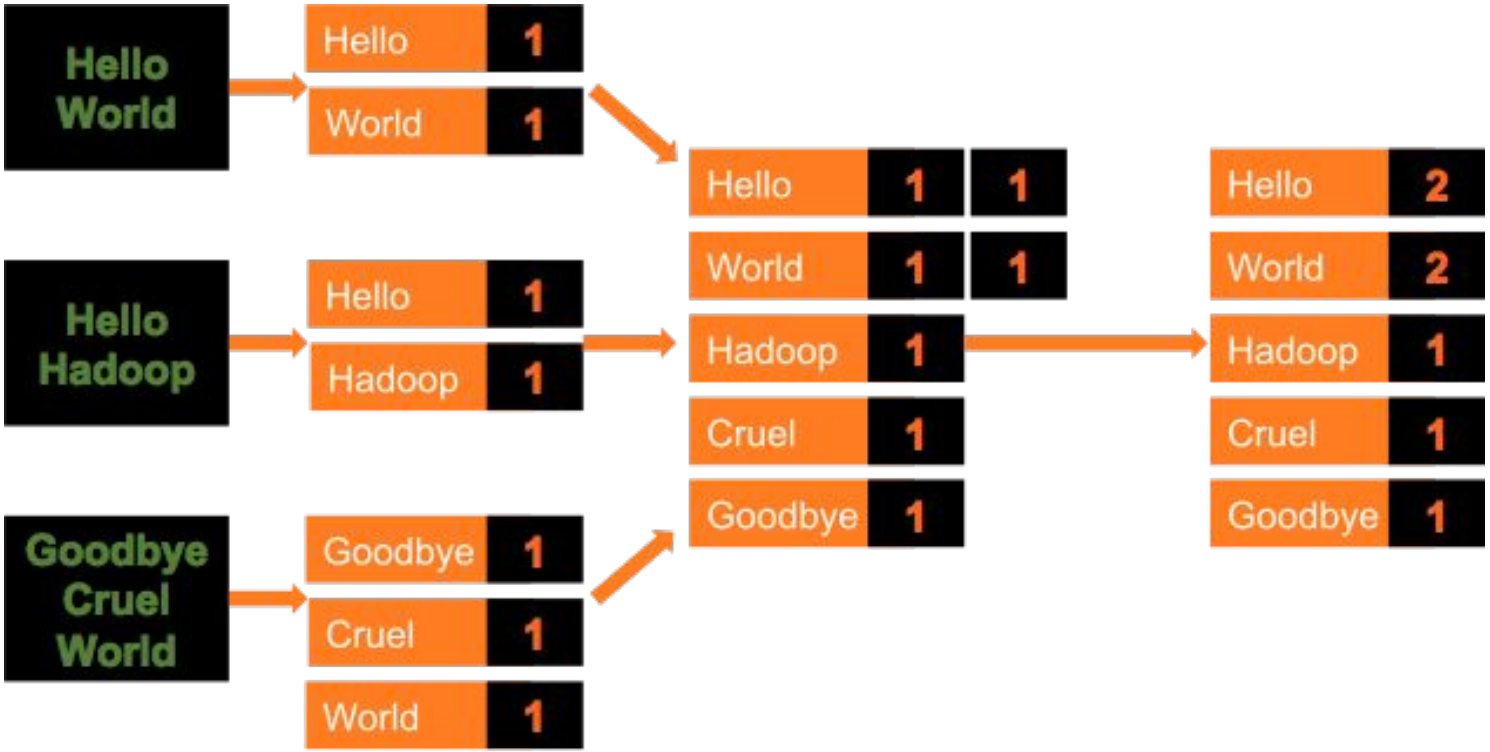
# Hadoop Core Components

- HDFS - Hadoop Distributed File System
  - Provides a distributed data storage system to store data in smaller blocks in a fail safe manner
- MapReduce - Programming framework
  - Has the ability to take a query over a dataset, divide it and run in in parallel on multiple nodes
- Yarn - (Yet Another Resource Negotiator) MRv2
  - Splitting a MapReduce Job Tracker's info
    - Resource Manager (Global)
    - Application Manager (Per application)

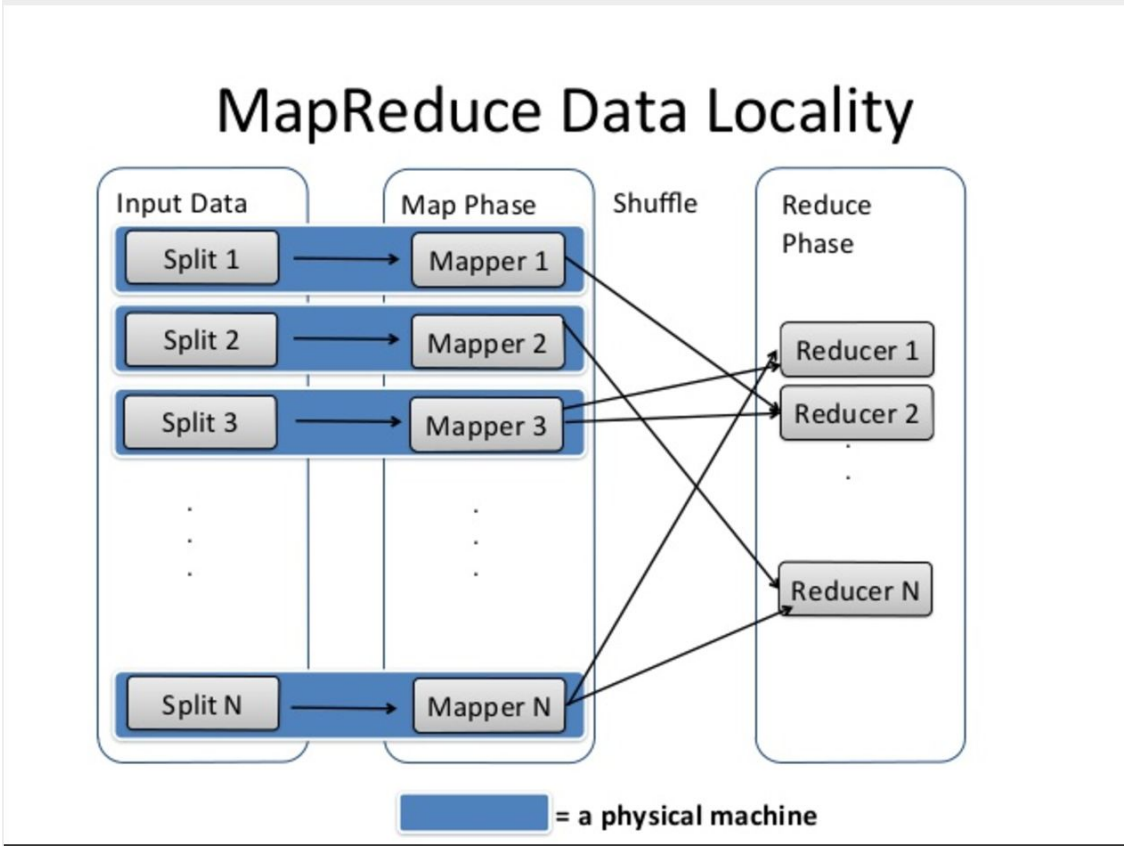



Panorays

# MapReduce via WordCount

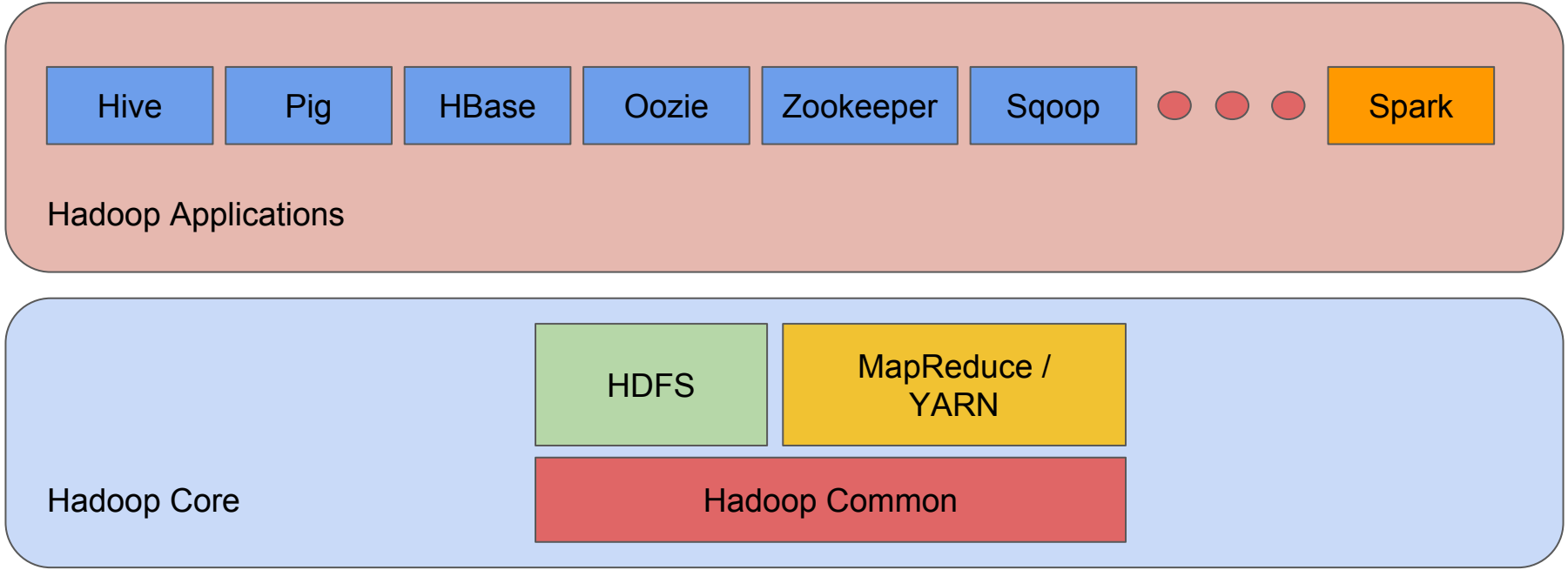


# Map/Reduce model and locality of data




Panorays

# Hadoop Ecosystem




Panorays

# Hadoop (+Spark) Distributions




Elastic MapReduce

DataProc

# What is ad hoc analysis on big data sets?

| Ad Hoc Anayltics | Relational Databases | Big Data Sets |
|---|---|---|
| Data Volume | Megabytes - Gigabytes | Terabytes - Petabytes |
| Data Velocity | Near real-time updates (Seconds) | Real-time updates (milliseconds) |
| Data Variety | Structured data | Structured and Unstructured Data |
| Data Model | 10s of tables/variables | 100s - 1000s of tables/variables |

https://www.qubole.com/resources/article/ad-hoc-analysis/

# New Age BI Applications

- Able to understand various types of data
- Ability to clean the data
- Process data with applied rules locally and in distributed environment
- Visualize sizeable data with speed
- Extend results by sharing within the enterprise

# Big Data Analytics

- Processing large amounts of data without data movement
- Avoid data connectors if possible (run natively)
- Ability to understand vast amount of data types and and data compressions
- Ability to process data on variety of processing frameworks
- Distributed data processing
  - In-Memory a big plus
- Super fast visualization
  - In-Memory a big plus

# Summary - When to choose hadoop?

- Large volumes of data to store and process
- Semi-Structured or Unstructured data
- Data is not well categorized
- Data contains a lot of redundancy
- Data arrives in streams or large batches
- Complex batch jobs arriving in parallel
- You don’t know how the data might be useful

Panorays


# Mind the Attack Surface