CPSC 436C

Cloud Computing for Data Science

Big Data

Maryam R.Aliabadi

mraiyata@cs.ubc.ca

Spring 2024

Last Week‘s Review

- Virtualization
- Virtualization types
- VM categories
- Partitioning
- VM Live migration
- How to launch a VM?

Configuring an instance based on the use case

Scenario
Name and tags
AMI
Instance type
Key pair
Network settings
Storage configurations
Advanced details

Your manager has asked you to create an EC2 instance that will host a dynamic website. After asking your manager more specific questions, you learn the following:

- The website should be available to everyone on the web, but the primary customer target is on the east coast of the United States.
- The instance that is hosting the website should have a Windows operating system.
- The application should be launched with the most recent patches and updates.
- The instance will need to have an administrator update patches from time to time.
- The instance must be protected from accidentally being terminated.
- The application will need to access Amazon S3.
- The instance's resources should be reportable for costs
- The costs to run the instance should be kept as low as possible.

Configuring Name and Tags

- Scenario
- Name and tags
- AMI
- Instance type
- Key pair
- Network settings
- Storage configurations
- Advanced details

Requirements to consider:

- The instance's resources should be reportable for costs.
- The costs to run the instance should be kept as low as possible.

To help track the costs of the instance, **tags** should be used. You can run reports based on tags to gain insight into the monthly costs of this particular instance.

Key	Value
Name	My test server
Dept	Development

Configuring AMI

- Scenario
- Name and tags
- AMI
- Instance type
- Key pair
- Network settings
- Storage configurations
- Advanced details

Requirements to consider:

- The instance that is hosting the website should have a Windows operating system.
- The costs to run the instance should be kept as low as possible.

An **AMI** should be chosen that is packaged with **Windows** as the operating system. Also, you should give careful consideration for any other software that might be needed when choosing your AMI. Your manager said to keep costs as low as possible. However, you don't want the performance of your website to suffer in order to keep costs low. Remember that AMIs cannot be switched out. If you later find that you need a more advanced AMI, you will need to create a new instance.

Amazon Machine Image components:

- **A template for the root volume**
- **Launch permissions**
- **A block device mapping**

Configuring Instance Type

- Scenario
- Name and tags
- AMI
- Instance type
- Key pair
- Network settings
- Storage configurations
- Advanced details

Requirements to consider:

- The website should be available to everyone on the web, but the primary customer target is on the east coast of the United States.
- The costs to run the instance should be kept as low as possible.

The instance is going to be used to host a website, and cost is a factor. Therefore, the most cost-effective instance type for a web server is one of the families in the **general purpose** category. Remember that instance types can be changed. You can scale up or scale down as needed. For example, you might start with an instance in the T3 family, and scale up if you need more CPU.

General purpose instances	
Instance types	A1, M4, M5, T2, T3
Use case	These instances are ideal for applications that use these resources in equal proportions such as **web servers** and code repositories.

Configuring Key Pair

- Scenario
- Name and tags
- AMI
- Instance type
- **Key pair**
- Network settings
- Storage configurations
- Advanced details

Requirements to consider:

- The instance will need to have an administrator update patches from time to time.

Because the instance will need to have an administrator update patches from time to time, the instance should be created with a key pair. You can either **create a new key pair or use an existing key pair**. The administrator who is making the updates should have access to this key pair.

A key pair consists of the following:

- A **public key** that AWS stores
- A **private key** file that you store

Configuring Network

- Scenario
- Name and tags
- AMI
- Instance type
- Key pair
- Network settings
- Storage configurations
- Advanced details

Requirements to consider:

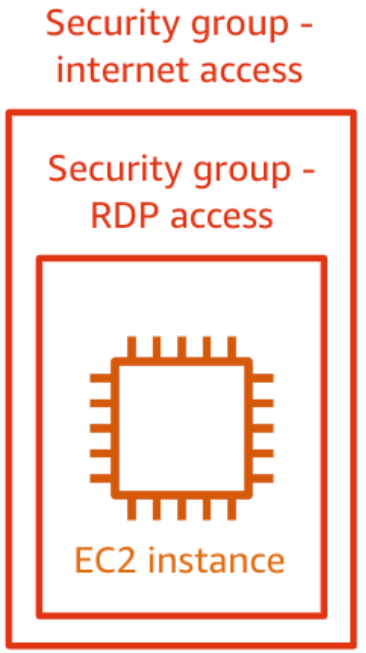
- The instance that is hosting the website should have a Windows operating system.
- The website should be available to everyone on the web, but the primary customer target is on the east coast of the United States.

The website will be primarily targeting customers on the east coast of the United States. Therefore, the best Region to launch your instance in is the **N. Virginia Region**. You should be sure that the **VPC** and **subnet** that you place your instance in are configured for internet access. Also, a **public IP address** should be assigned to the instance.

The **security group** should have inbound rules that allow for the following:

- Internet (HTTP/HTTPS) traffic for the public website to be accessible to the internet.

- Remote desktop protocol (RDP) traffic for an administrator to log in for patching and updating the instance. If the instance had a Linux OS, then you would use SSH instead.



Configuring Storage

- Scenario
- Name and tags
- AMI
- Instance type
- Key pair
- Network settings
- **Storage configurations**
- Advanced details

Requirements to consider:

- The website should be available to everyone on the web, but the primary customer target is on the east coast of the United States.
- The costs to run the instance should be kept as low as possible.

A **general purpose EBS volume** would make the best choice for this workload. A provisioned IOPS volume is over-resourcing and will cost you more than you need to spend. If the website was hosting a critical business website with a large database, then a provisioned IOPS volume could be the right solution. You can always scale to meet the needs of your storage.

Amazon EBS capabilities:

Run databases

Host applications

Handle most storage computing needs

Configuring Access Control

- Scenario
- Name and tags
- AMI
- Instance type
- Key pair
- Network settings
- Storage configurations
- Advanced details

Requirements to consider:

- The instance must have an administrator update patches from time to time.
- The application must be protected from accidentally being terminated.
- The application will need to access Amazon S3

Because the application must access Amazon S3, you should attach an **IAM role** to the instance that has sufficient permissions to perform the required tasks.

EC2 instance with an IAM role

S3 bucket

To protect the instance from accidental termination, enable **termination protection**.

To update and patch the instance when it is launched, add the appropriate script to the **user data** field.

Module 1 Quiz

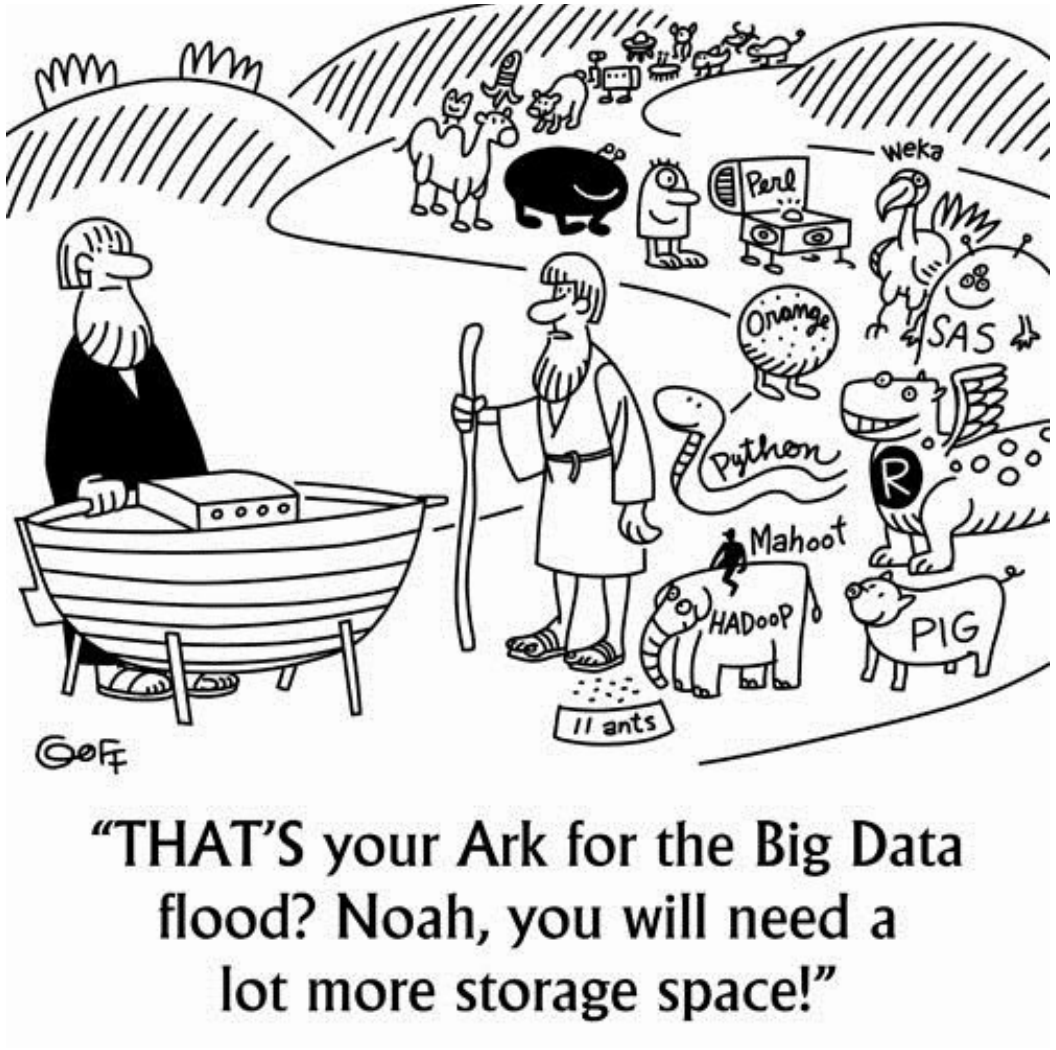
- https://join.iclicker.com/YFMA

Module 1 Survey

- https://docs.google.com/forms/d/e/1FAIpQLScz9UH1ktXcmSed1pzLsxFKiB7xeuefJqhfyYe2ptk38S_87w/viewform?usp=sf_link

Today's Topics

- Big data definition
- Big data properties
- Big data sources
- Big data analytics stack



"THAT'S your Ark for the Big Data flood? Noah, you will need a lot more storage space!"

Big Data

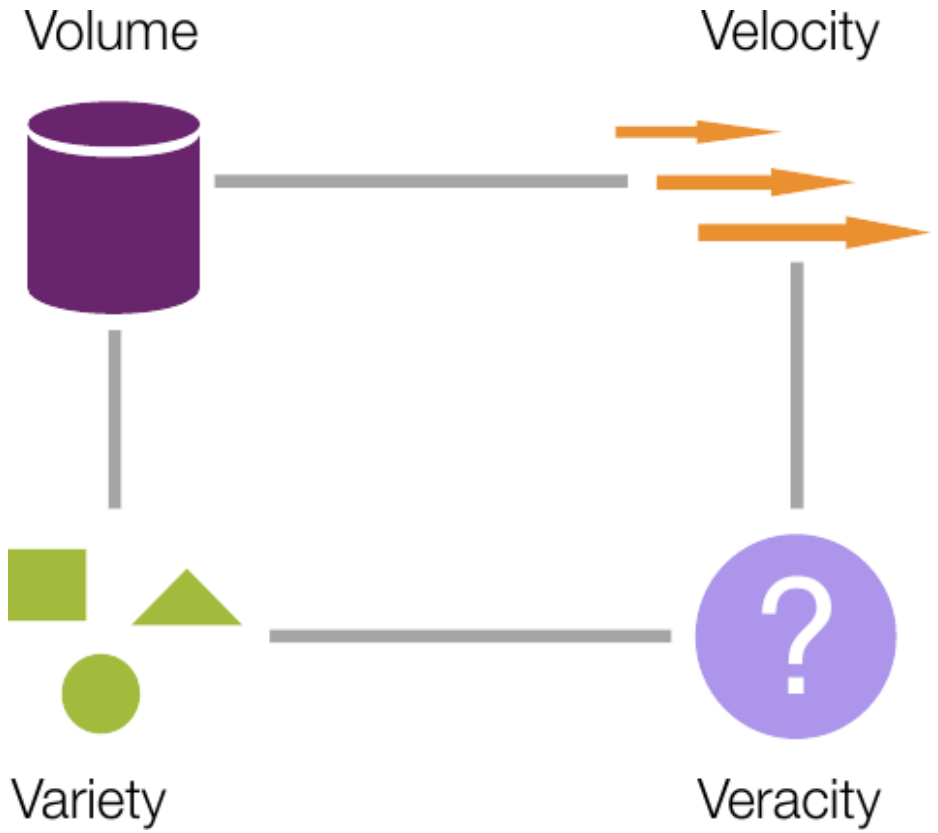
- Big Data refers to datasets and flows large enough that has outpaced our capability to store, process, analyze, and understand.

small data

big data

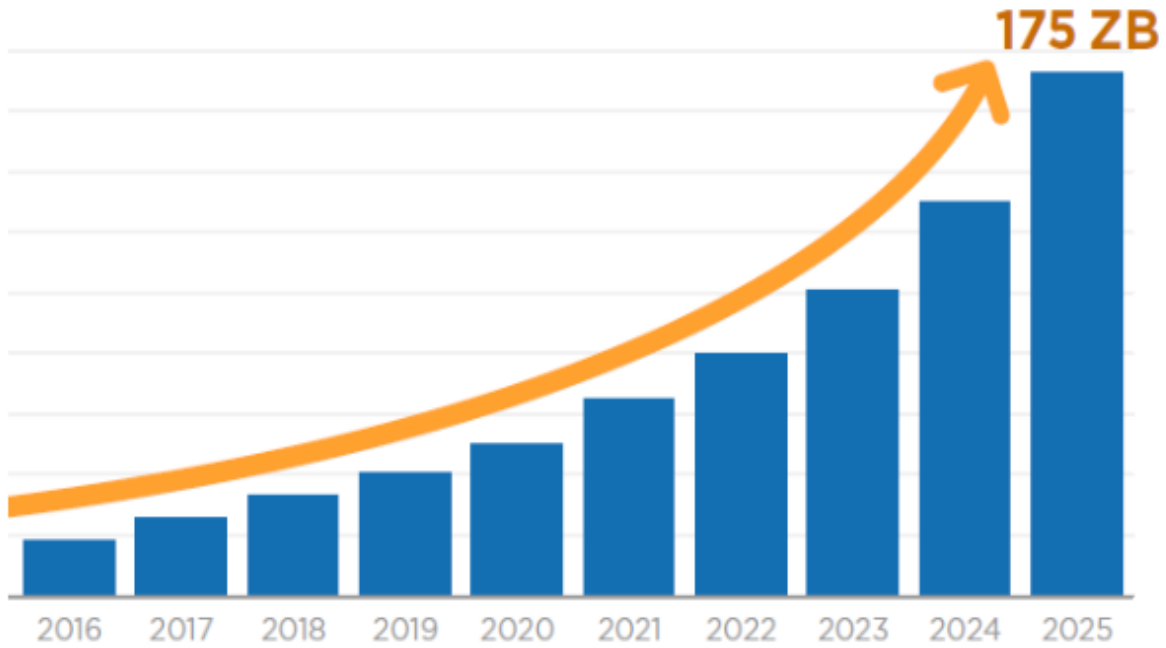
Four Attributes of Big Data

- Volume: data size
- Velocity: data generation rate
- Variety: data heterogeneity
- Veracity: data quality



Volume

- More data than fits in a computer's RAM
- More data than fits on a single hard drive
- Facebook's 2+billion users



Velocity

- Rapid generation of new data
- Nearly 200 million emails are sent each minute of each day
- Nearly 5 billion videos are watched on YouTube every day

Variety

- Data in many formats
- Videos, photos, audio
- GPS coordinates
- Social network connections

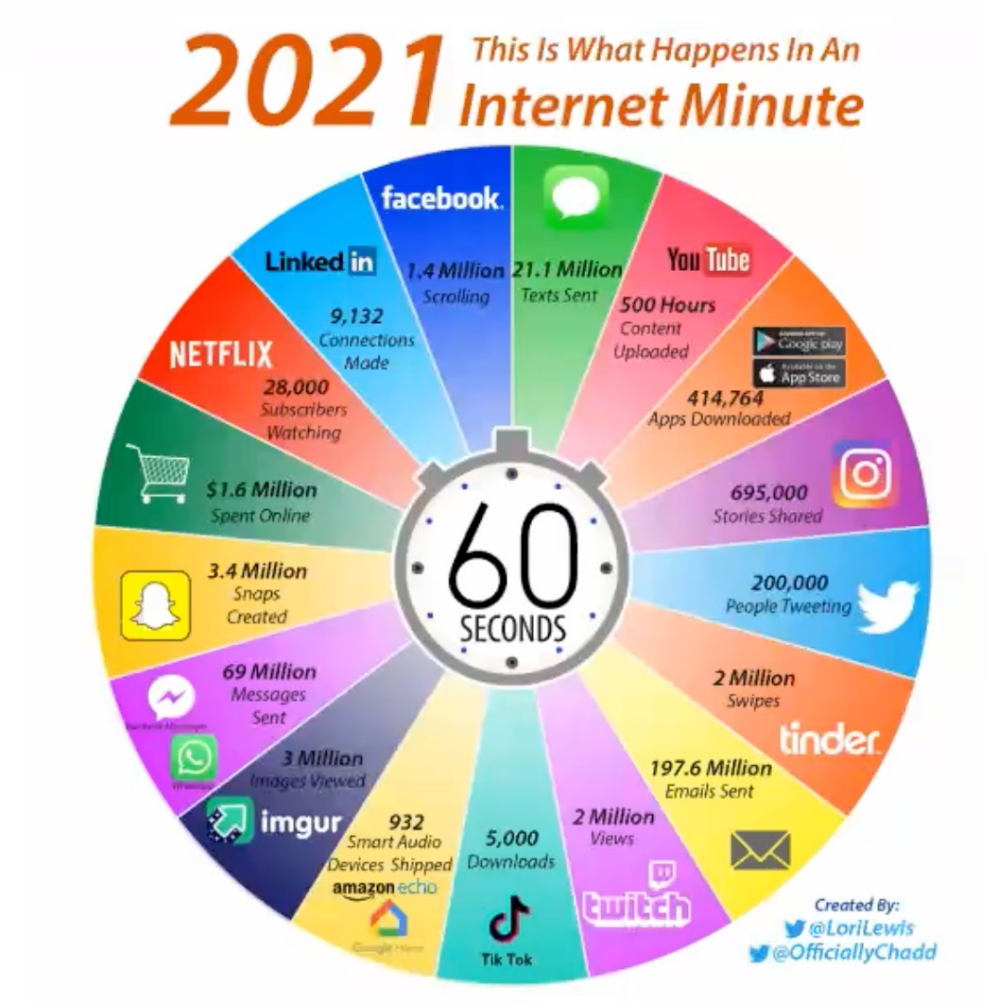
Veracity

- Data reliability and trustworthiness
- Important to make informed decisions or draw meaningful insights.
- Data quality processes, data validation procedures, and data governance practices are required to maintain and improve the accuracy and trustworthiness of their data.

Where does big data come from?

Big Data Market Driving Factors

- Social Media
  - Social media begets more social media
  - Posts get liked
  - Images get tags
  - Followers share content



Big Data Market Driving Factors

- Internet of Things (IoT)
  - IoT sensors
  - Smart homes
  - Smart grids
  - Self-driving cars

- More than 65 billion devices were connected to the Internet by 2010, and this number exceeded 230 billion by 2020.

How to store and process big data?

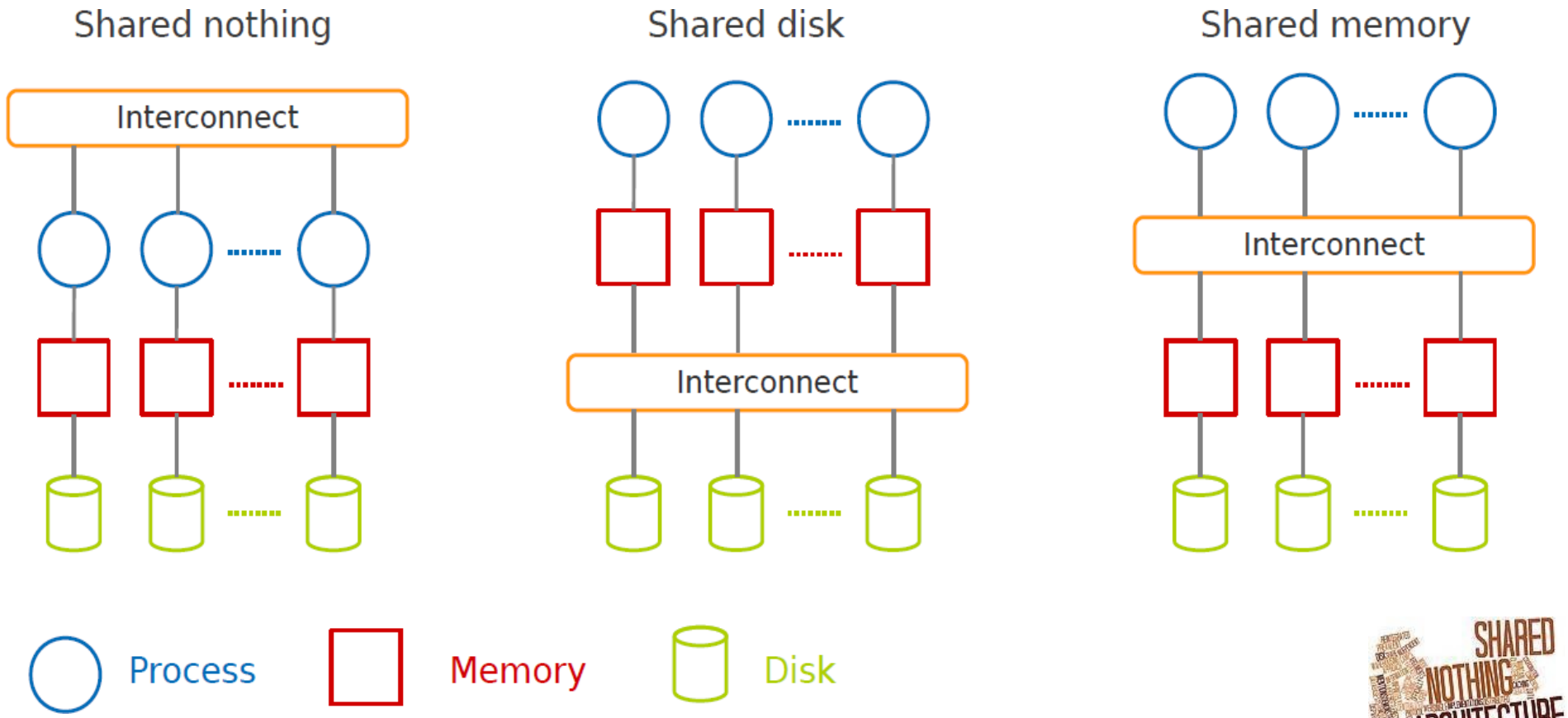
Scale Up vs. Scale Out

- Scale up or scale vertically: adding resources to a single node in a system.
- Scale out or scale horizontally: adding more nodes to a system.

Scale Up vs. Scale Out

- Scale up: more expensive than scaling out.
- Scale out: more challenging for fault tolerance and software development.

Taxonomy of Parallel Architectures

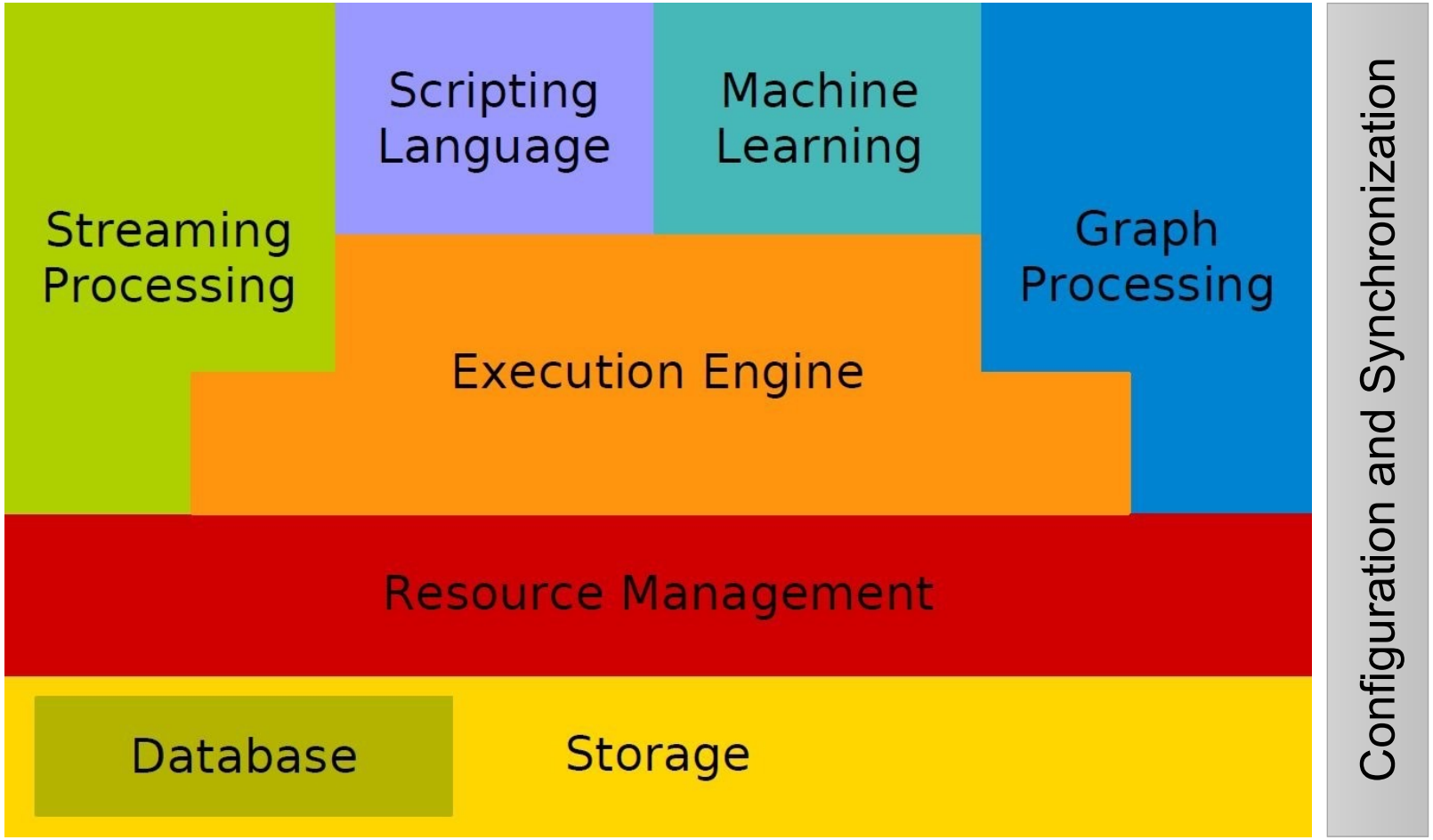


DeWitt, D. and Gray, J. "Parallel database systems: the future of high performance database systems". ACM Communications, 35(6), 85-98, 1992.

Big Data Tools and Frameworks

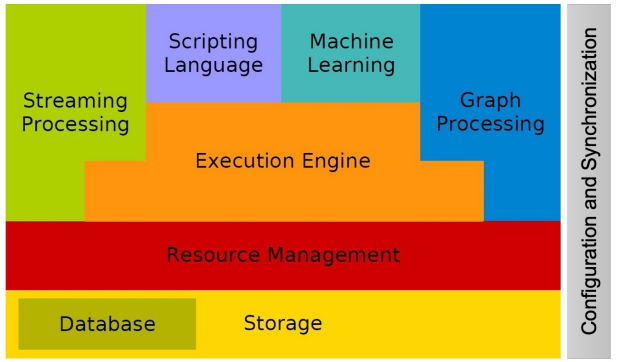
- Two main types of tools:
  - Data Store
  - Data Processing

Big Data Analytics Stack



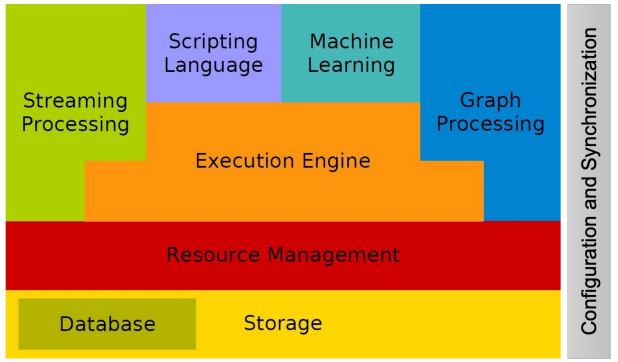
Big Data – Storage (File systems)

- Traditional file-systems 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/GFS, Amazon S3, ...



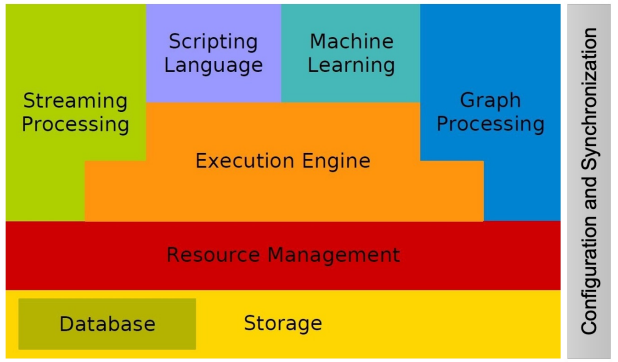
Big Data - Database

- Relational Databases Management Systems (RDMS) were not designed to be distributed.
- NoSQL databases relax one or more of the ACID properties:
  - BASE (Basically Available, Soft state, Eventually consistent).
- Different data models: key/value, column-family, graph, document.
- NoSQL database exeamples: Hbase/BigTable, Dynamo, Scalaris, Cassandra, MongoDB, Voldemort, Riak, Neo4J, ...



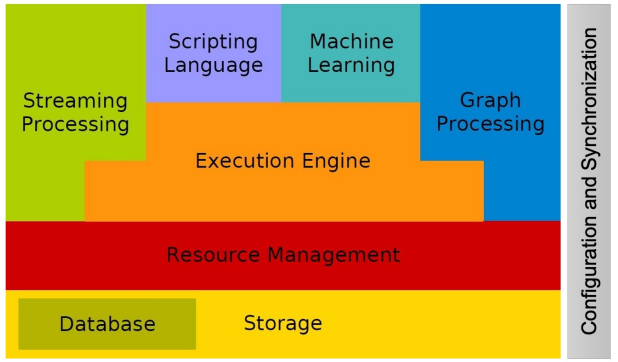
Big Data – Resource Management

- 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, Borg, …



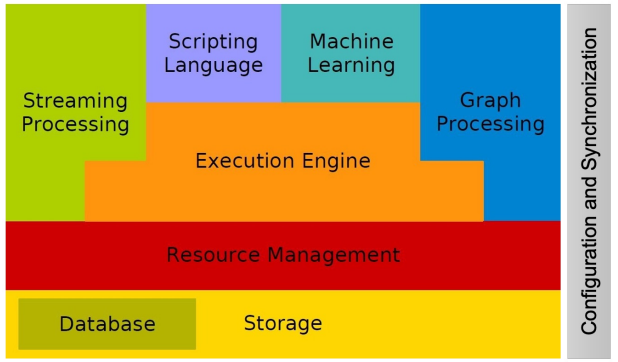
Big Data- Execution Engines

- 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, ...



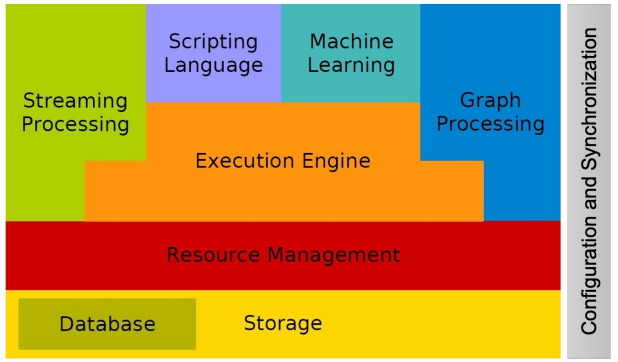
Big Data – Query/Scripting Languages

- 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 execution engines.
- It translates user-defined functions to low-level API of the execution engines.
- Pig, Hive, Shark, Meteor, DryadLINQ, SCOPE, ...



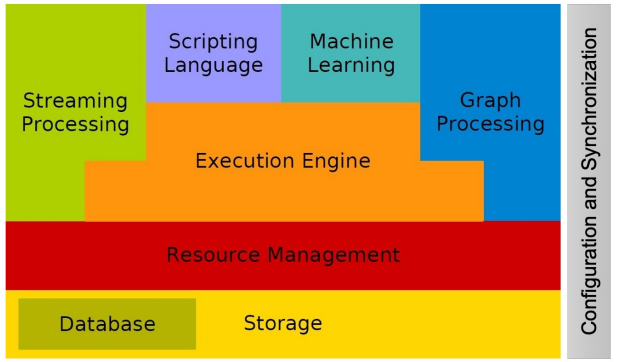
Big Data: Graph Processing

- 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 problems.
- Pregel, Giraph, GraphX, GraphLab, PowerGraph, GraphChi, …



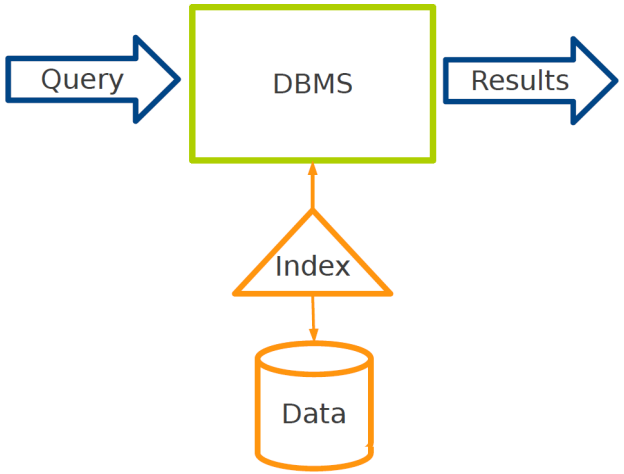
Big Data – Machine Learning

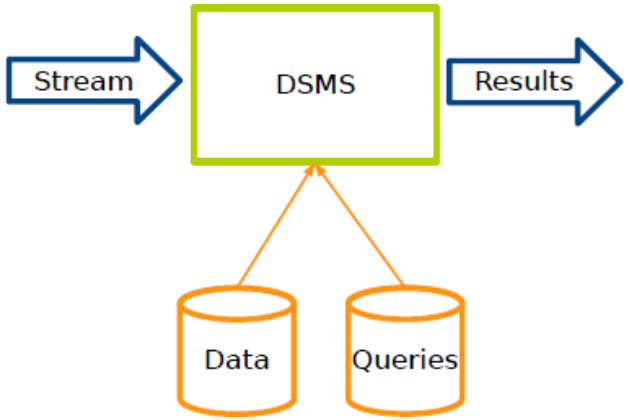
- Implementing and consuming machine learning techniques at scale are difficult tasks for developers and end users.
- There exist platforms that address it by providing scalable machine-learning and data mining libraries.
- Mahout, MLBase, Tensorflow, ...

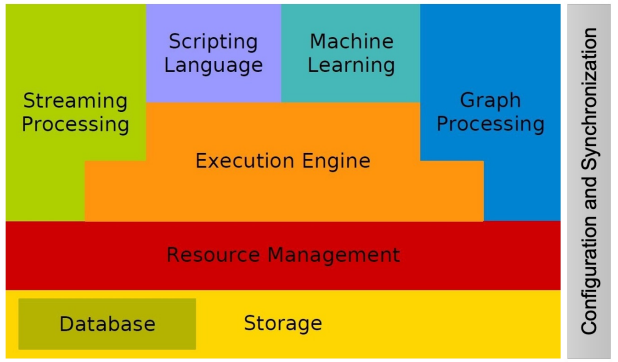


Big Data – Stream Processing

- Providing users with fresh and low latency results.
- Database Management Systems (DBMS) vs. Data Stream Management Systems (DSMS)
- Storm, S4, SEEP, D-Stream, Naiad, ...

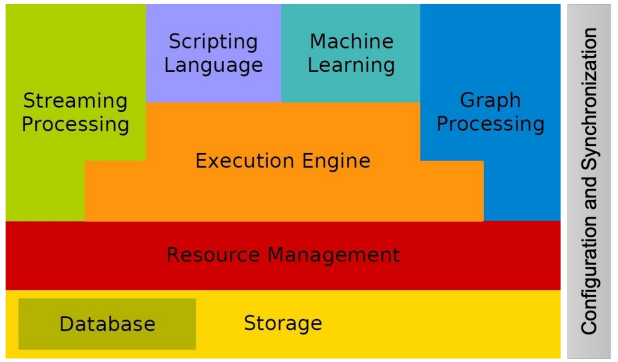






Big Data – Configuration and Synchronization

- A means to synchronize distributed applications accesses to shared resources.
- Allows distributed processes to coordinate with each other.
- Zookeeper, Chubby, ...



Recap

- Big data definition
- Big data properties
- Big data sources
- Big data analytics stack

Next Topic: Data Store