State of Art:
Big Data Technologies

Big Data - MSL7300

IIT Jodhpur

Revendranath T

Outline

1. An E-Commerce case example
2. Challenges with big data
3. How to solve big data challenges?
4. Brief history of big data
5. Two big data technologies
    - Snowflake
    - Databricks
6. Next steps

KhanMarket.com: What is your approach?

- Suppose **KhanMarket.com** is an e-commerce company that sells Indian made goods to consumers in Indian market
- Your goal is to predict purchases behaviour of an user based on browsing.
  - **Objective**: What is the probability that a user actually buys something from your store in a given session?
  - **Assumption**: 100k active daily users, 2 million overall registered users

KhanMarket.com: What is your approach?

- Suppose **KhanMarket.com** is an e-commerce company that sells Indian made goods to consumers in Indian market
- Your goal is to predict purchases behaviour of an user based on browsing.
  - **Objective**: What is the probability that a user actually buys something from your store in a given session?
  - **Assumption**: 100k active daily users, 2 million overall registered users
- Approach:
  - Store data from the web/mobile etc.
  - Import/export data from storage
  - Filter/clean data
  - Model, valide, deploy, visualize, feedback etc.

KhanMarket.com: What is your approach?

- Suppose **KhanMarket.com** is an e-commerce company that sells Indian made goods to consumers in Indian market
- Your goal is to predict purchases behaviour of an user based on browsing.
  - **Objective**: What is the probability that a user actually buys something from your store in a given session?
  - **Assumption**: 100k active daily users, 2 million overall registered users
- Approach:
  - Store data from the web/mobile etc.
  - Import/export data from storage
  - Filter/clean data
  - Model, valide, deploy, visualize, feedback etc.
- The data continues to grow with:
  - usage of current users, and addition of new users
- **Challenges?**

Big Data: Challenges

1. Difficult to handle
2. Hard to extract value from big data due to size and complexity

Big Data: Challenges

1. Difficult to handle
    a. Unorthodox data sources
    b. Poorly structured data
        i. Raw text, web pages, images etc
    c. Infrastructure to handle varieties of large data
2. Hard to extract value from big data due to size and complexity
    a. Computational, methodology & models challenges to extract insights from large volumes of data
    b. **bigP** problem:
        i. A dataset has close to or even more variables (columns) than observations
    c. **bigN** problem:
        i. A dataset has massive numbers of observations (rows) such that it cannot be handled with standard data analytics techniques and/or on a standard desktop computer.

Big Data: Challenges: How do you solve?

1. Difficult to handle
    a. Unorthodox data sources
    b. Poorly structured data
    c. Infrastructure to handle varieties of large data
2. Hard to extract value from big data due to size and complexity
    a. Computational, methodology & models
    b. bigP problem:
    c. bigN problem:

bigP Problem: Review for KhanMarket.com

- Suppose **KhanMarket.com** is an e-commerce company that sells Indian made goods to consumers in Indian market
- Your goal is to predict purchases behaviour of an user based on browsing.
  - **Objective**: What is the probability that a user actually buys something from your store in a given session?
  - **Assumption**: 100k active daily users, 2 million overall registered users
  - **Dependent variable**: Purchased (1: Yes, 2: No)
  - **Independent variable**: Source of visit (Youtube, Google, Bing, Twitter, Facebook, etc)
- Model: Thoughts?

bigP Problem: Review for KhanMarket.com

- Suppose **KhanMarket.com** is an e-commerce company that sells Indian made goods to consumers in Indian market
- Your goal is to predict purchases behaviour of an user based on browsing.
  - **Objective**: What is the probability that a user actually buys something from your store in a given session?
  - **Assumption**: 100k active daily users, 2 million overall registered users
  - **Dependent variable**: Purchased (1: Yes, 2: No)
  - **Independent variable**: Source of visit (Youtube, Google, Bing, Twitter, Facebook, etc)
- Model: Logistic Regression (or any binary classifier)

bigP Problem: Review for KhanMarket.com

- **Dependent variable**: Purchased (1: Yes, 2: No)
- **Independent variable**: Source of visit (Youtube, Google, Bing, Twitter, Facebook, etc)
- Model: Logistic Regression (or any binary classifier)

	Estimate	Pr(>|z|)
(Intercept)	-1.3831	0.000e+00
bing	-1.4647	4.416e-03
dfa	-0.1865	1.271e-01
docs.google.com	-2.0181	4.714e-02
facebook.com	-1.1663	3.873e-04
google	-1.0149	6.321e-168
google.com	-2.9607	3.193e-05
m.facebook.com	-3.6920	2.331e-04
Partners	-4.3747	3.942e-14
quora.com	-3.1277	1.869e-03
siliconvalley.about.com	-2.2456	1.242e-04
sites.google.com	-0.5968	1.356e-03
t.co	-2.0509	4.316e-03
youtube.com	-6.9935	4.197e-23

Observations?

bigP Problem: Review for KhanMarket.com

- **Dependent variable**: Purchased (1: Yes, 2: No)
- **Independent variable**: Source of visit (Youtube, Google, Bing, Twitter, Facebook, etc)
- Model: Logistic Regression (or any binary classifier)

	Estimate	Pr(>|z|)
(Intercept)	-1.3831	0.000e+00
bing	-1.4647	4.416e-03
dfa	-0.1865	1.271e-01
docs.google.com	-2.0181	4.714e-02
facebook.com	-1.1663	3.873e-04
google	-1.0149	6.321e-168
google.com	-2.9607	3.193e-05
m.facebook.com	-3.6920	2.331e-04
Partners	-4.3747	3.942e-14
quora.com	-3.1277	1.869e-03
siliconvalley.about.com	-2.2456	1.242e-04
sites.google.com	-0.5968	1.356e-03
t.co	-2.0509	4.316e-03
youtube.com	-6.9935	4.197e-23

Observations

1. Reject all null hypothesis?
2. Not trust the coefficient t- tests
3. Lack of predictive ability of the model

How do you change the model?

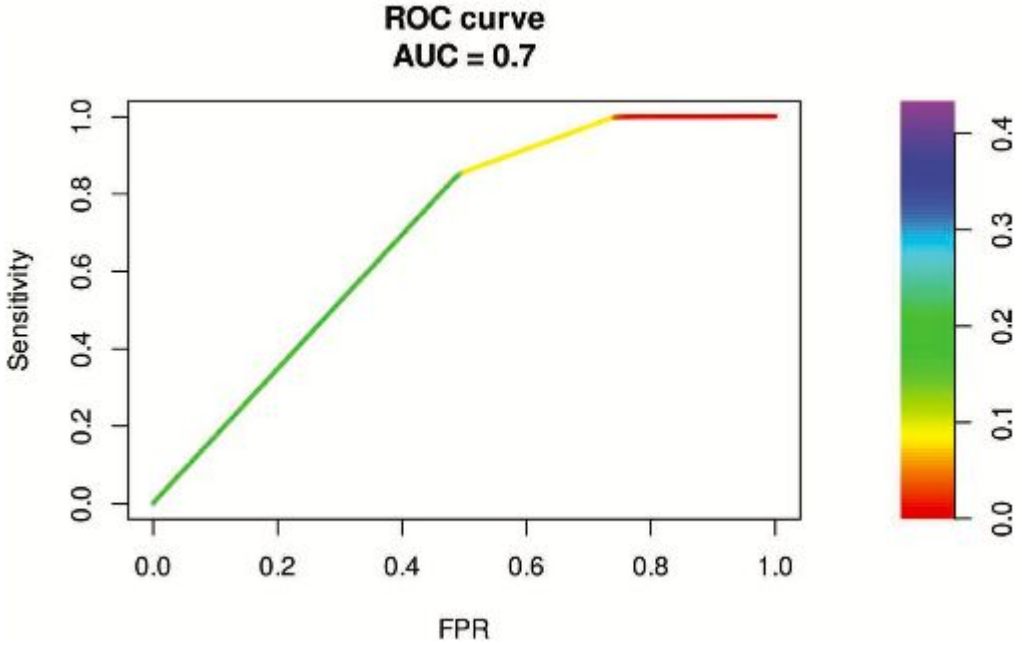
1. Add or remove variables
2. Look at co-variance & select the variables
3. Find possible combinations of co-variates & build several models. Select relevant model.
    a. Computational challenges
    b. Personal biases

bigP Problem: Review for KhanMarket.com

- **Dependent variable**: Purchased (1: Yes, 2: No)
- **Independent variable**: Source of visit (Youtube, Google, Bing, Twitter, Facebook, etc)
- Model: Logistic Regression (or any binary classifier)

Regularization: the lasso estimator

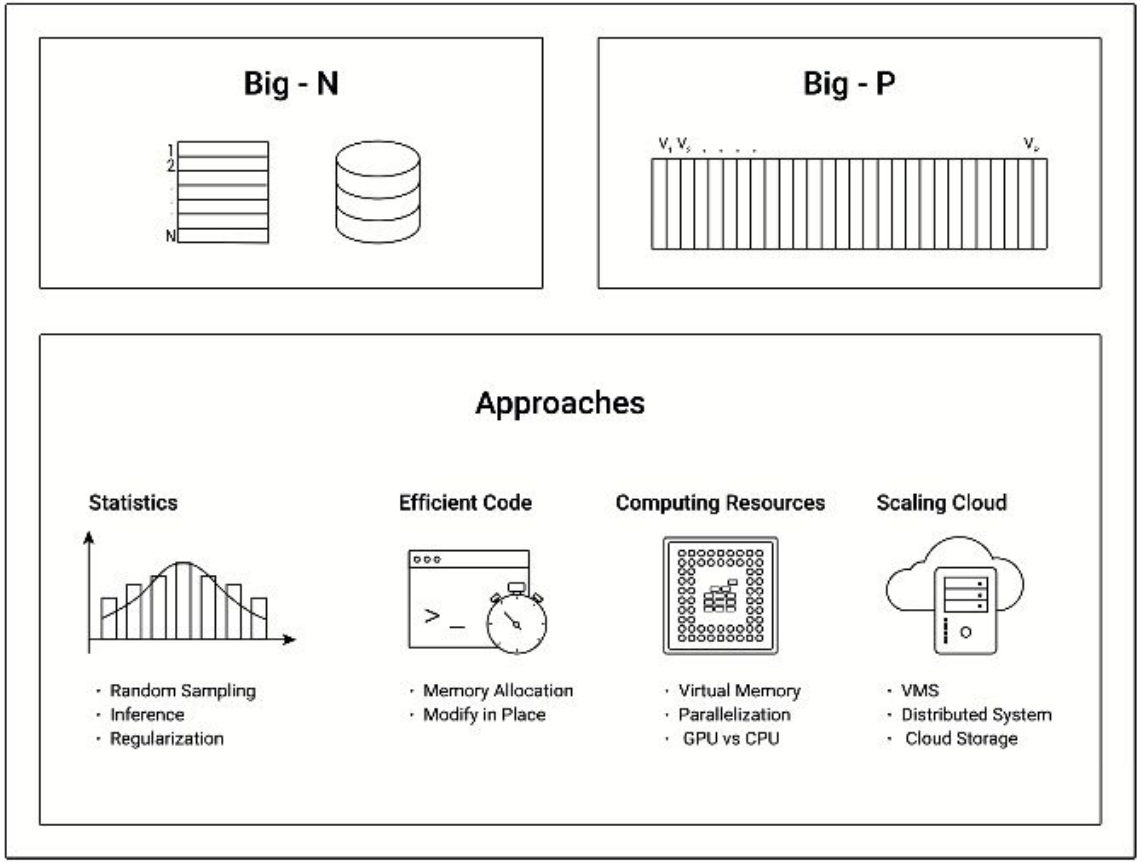
1. Convenient and efficient way to get a sequence of candidate models
2. Penalizes model complexity (the cause of instability) during the estimation procedure.



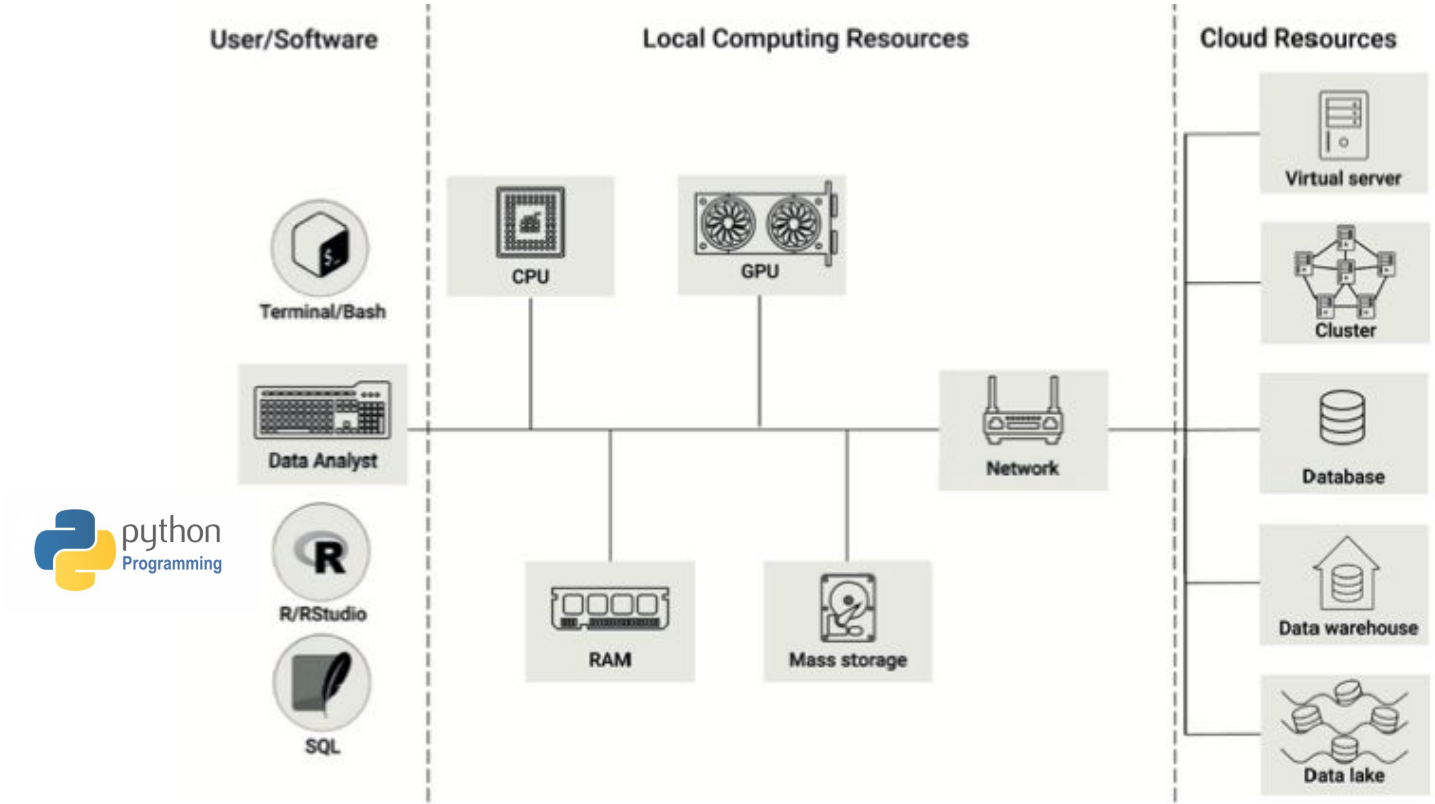
bigN Problems: Size larger than columns

1. Extract subset of the overall data and conduct analysis
2. Speed up the analysis on whole dataset using Uluru algorithm or others that are proven to reduce computational overhead

Four Perspectives to Solve BigP or BigN problems



Infrastructure for Big Data

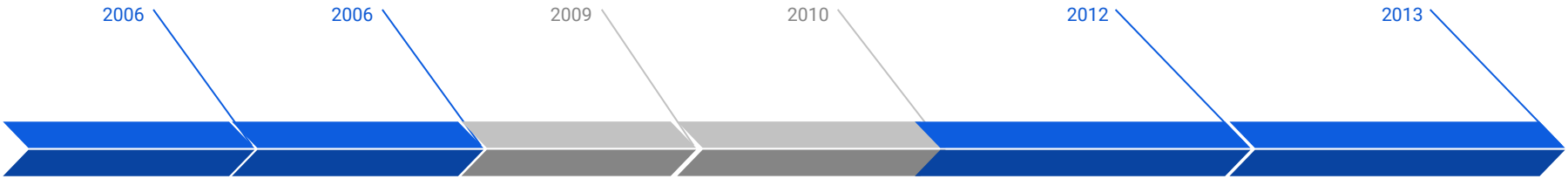


Measure Performance of Code Efficiency

package	function	purpose
`utils`	`object.size()`	Provides an estimate of the memory that is being used to store an R object.
`pryr`	`object_size()`	Works similarly to `object.size()`, but counts more accurately and includes the size of environments.
`pryr`	`mem_used()`	Returns the total amount of memory (in megabytes) currently used by R.
`pryr`	`mem_change()`	Shows the change in memory (in megabytes) before and after running code.
`base`	`system.time()`	Returns CPU (and other) times that an R expression used.
`microbenchmark`	`microbenchmark()`	Highly accurate timing of R expression evaluation.
`bench`	`mark()`	Benchmark a series of functions.
`profvis`	`profvis()`	Profiles an R expression and visualizes the profiling data (usage of memory, time elapsed, etc.).

Package/Method	Purpose
timeit.timeit	measuring the execution time
cProfile	measuring the execution time of large code blocks
profile.memory_profiler	monitoring memory consumption of a process as well as line-by-line analysis of memory consumption
perfplot	compare the performance of different functions

Big Data: History



Amazon S3

Amazon introduced S3 as a part of offering to AWS

Store files on S3

Hadoop

Apache Software Foundation released Hadoop

Expensive & Complex due to intense coding knowledge

Spark

University of California project

Open sourced in 2010

2012-23: widely adopted across the industry & a leading big data technology

Big Query

Announced by Google in 2010 & made available in 2011

2012-21 Support for SQL, Geospatial data, streaming data, ML/AI models

Easy to store & analyse big data at low cost

Snowflake

Founded in 2012 & launched in 2014

Data platform as a managed self-service

Cloud infrastructure made easy to store structured, semi-structured & unstructured data

SQL driven & easy for users to begin with Snowflake

DataBricks

Founded by original creators of Spark

2021: Publicly listed on NYSE

Made it easy for businesses to build & run big data applications

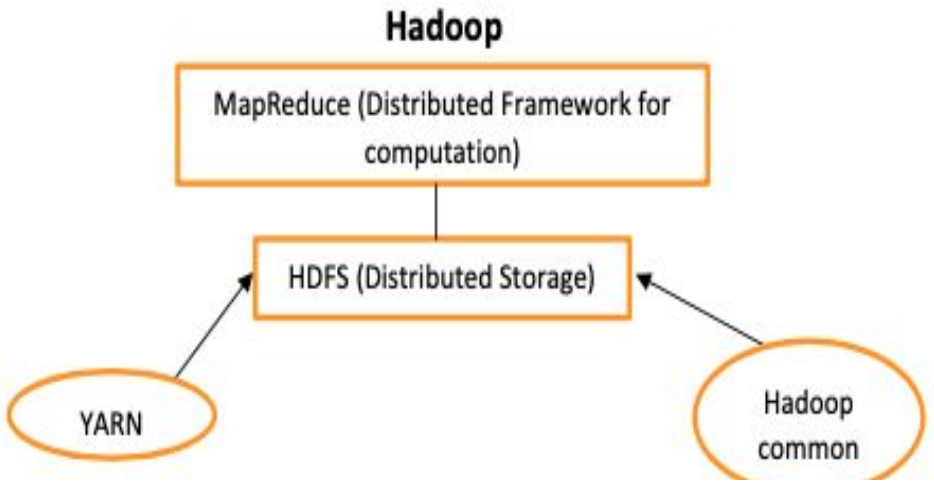
Map Reduce

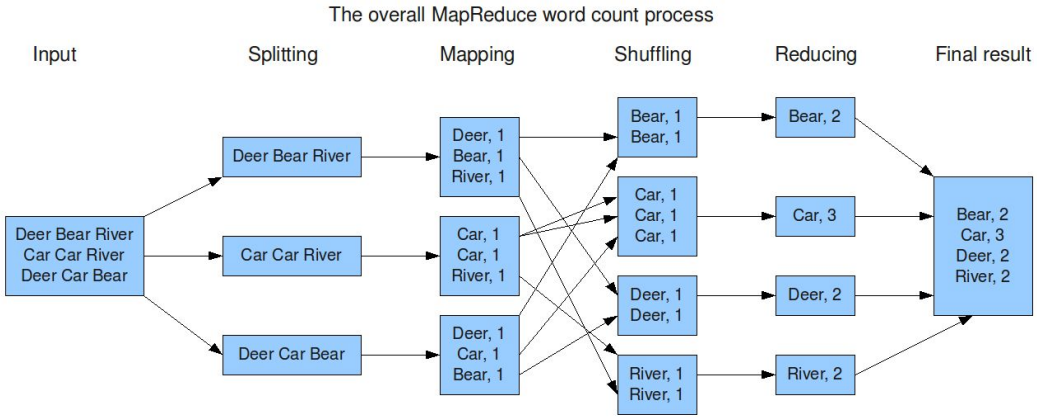
on)
data

e carries

luster
Data on

similar
ine”





Spark

- Spark is a fast and general-purpose cluster computing system.
- Uses
  - Hadoop Distributed Files System to store data
  - YARN to manage resources for distributed data
- Spark complements Hadoop

Key features

• **Batch processing**: Process large datasets in a batch-oriented manner (Similar to how Hadoop MapReduce works)
• **Streaming**: Process streaming data: data is constantly being generated, such as data from social media or sensors.
• **Machine learning**: Spark can be used to train and deploy machine learning models
• **Graph processing**: Process graphs to represent data as a network of nodes and edges.

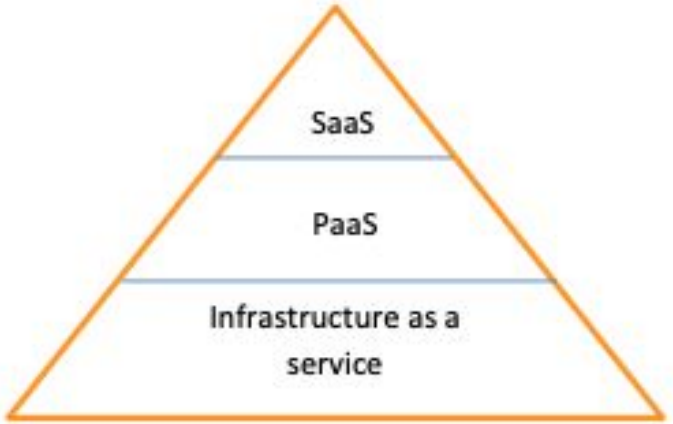
Spark & Hadoop: Comparison

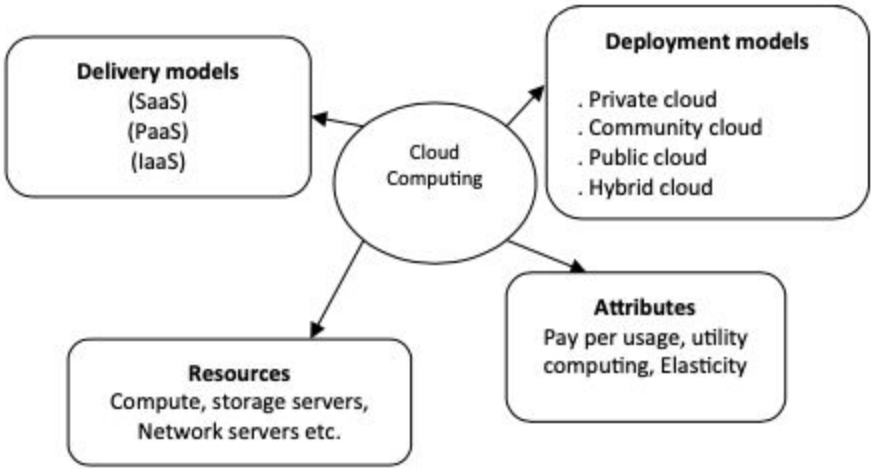
Features Comparison		
Feature	**Spark**	**Hadoop**
Programming model	Resilient Distributed Datasets (RDDs)	MapReduce
Speed	Faster for certain types of tasks	Slower for certain types of tasks
Scalability	Scalable to large clusters	Scalable to large clusters
Fault tolerance	Fault-tolerant	Fault-tolerant
Open source	Yes	Yes

Performance Comparison		
Feature	**Hadoop**	**Spark**
Batch processing	Better for large datasets	Worse for large datasets
Interactive queries	Worse for interactive queries	Better for interactive queries
Streaming data	Worse for streaming data	Better for streaming data
Memory usage	Less memory-efficient	More memory-efficient
Ease of use	More complex to use	Easier to use

Big Data as a Cloud Service

- AWS
- Azure
- Google Cloud Platform
- Databricks
- Snowflake
- Big Query





databricks

- Databricks is a unified, open analytics platform for building, deploying, sharing, and maintaining enterprise-grade data, analytics, and AI solutions at scale.
- The Databricks Lakehouse Platform integrates with cloud storage and security in your cloud account, and manages and deploys cloud infrastructure on your behalf.
- Uses of Databricks:
    - to process, store, clean, share, analyze, model, and monetize their datasets with solutions from BI to machine learning.
    - to build and deploy data engineering workflows, machine learning models, analytics dashboards, and more.

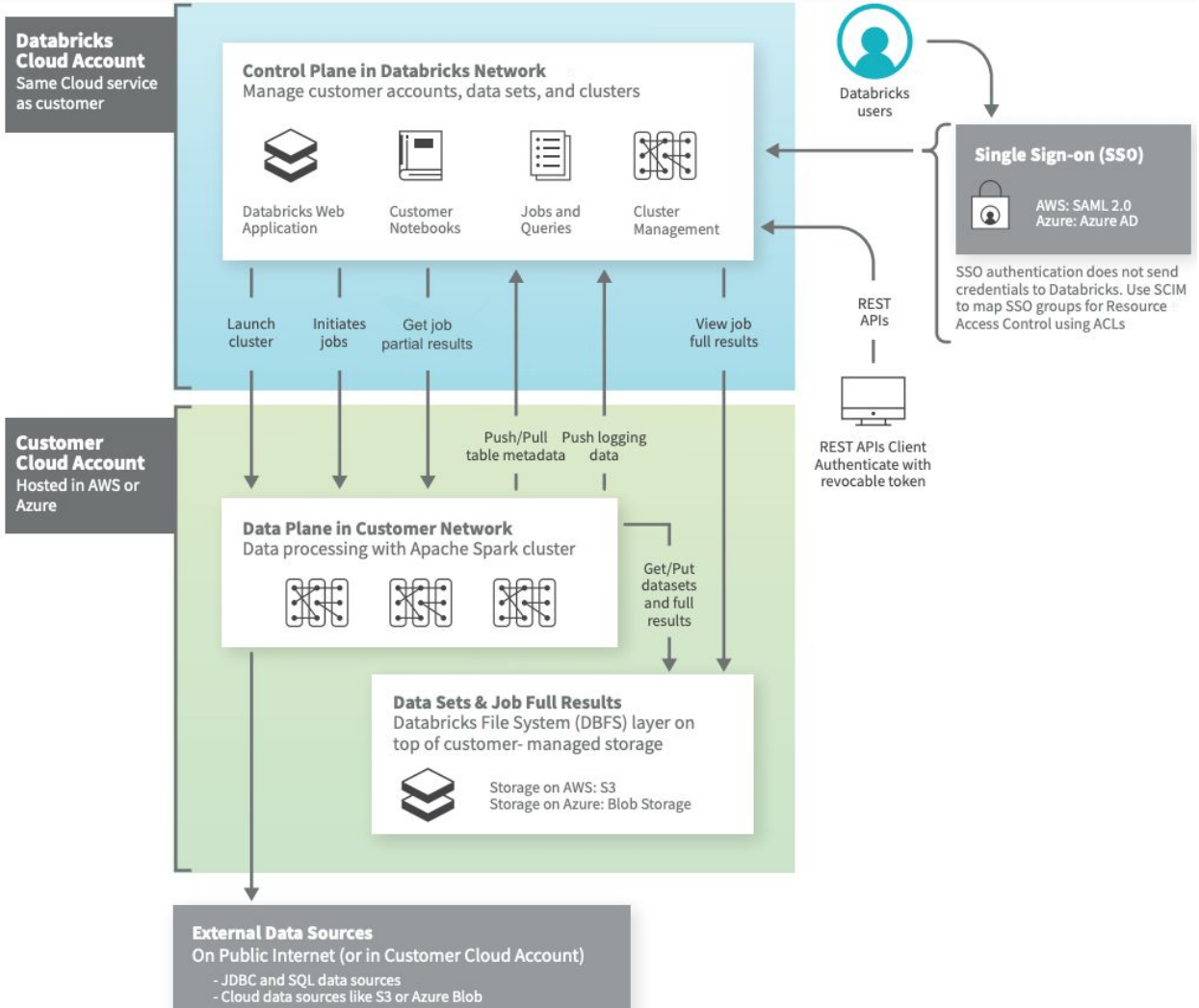
databricks

Databricks Workspace

- Data processing workflows scheduling and management
- Working in SQL
- Generating dashboards and visualizations
- Data ingestion
- Managing security, governance, and HA/DR
- Data discovery, annotation, and exploration
- Compute management
- Machine learning (ML) modeling and tracking
- ML model serving
- Source control with Git

- Integrates/Provides
  - Open source platforms
  - Command line interface
  - REST APIs
  - IDEs
  - Other BI tools
  - Git version control

databricks


Databricks Cloud Account
Same Cloud service as customer
Control Plane in Databricks Network
Manage customer accounts, data sets, and clusters
Databricks Web Application
Customer Notebooks
Jobs and Queries
Cluster Management
Databricks users
Single Sign-on (SSO)
AWS: SAML 2.0
Azure: Azure AD
SSO authentication does not send credentials to Databricks. Use SCIM to map SSO groups for Resource Access Control using ACLs
REST APIs
REST APIs Client Authenticate with revocable token
Launch cluster
Initiates jobs
Get job partial results
View job full results
Push/Pull table metadata
Push logging data
Customer Cloud Account
Hosted in AWS or Azure
Data Plane in Customer Network
Data processing with Apache Spark cluster
Get/Put datasets and full results
Data Sets & Job Full Results
Databricks File System (DBFS) layer on top of customer- managed storage
Storage on AWS: S3
Storage on Azure: Blob Storage
External Data Sources
On Public Internet (or in Customer Cloud Account)
- JDBC and SQL data sources
- Cloud data sources like S3 or Azure Blob

Snowflake's Data Cloud

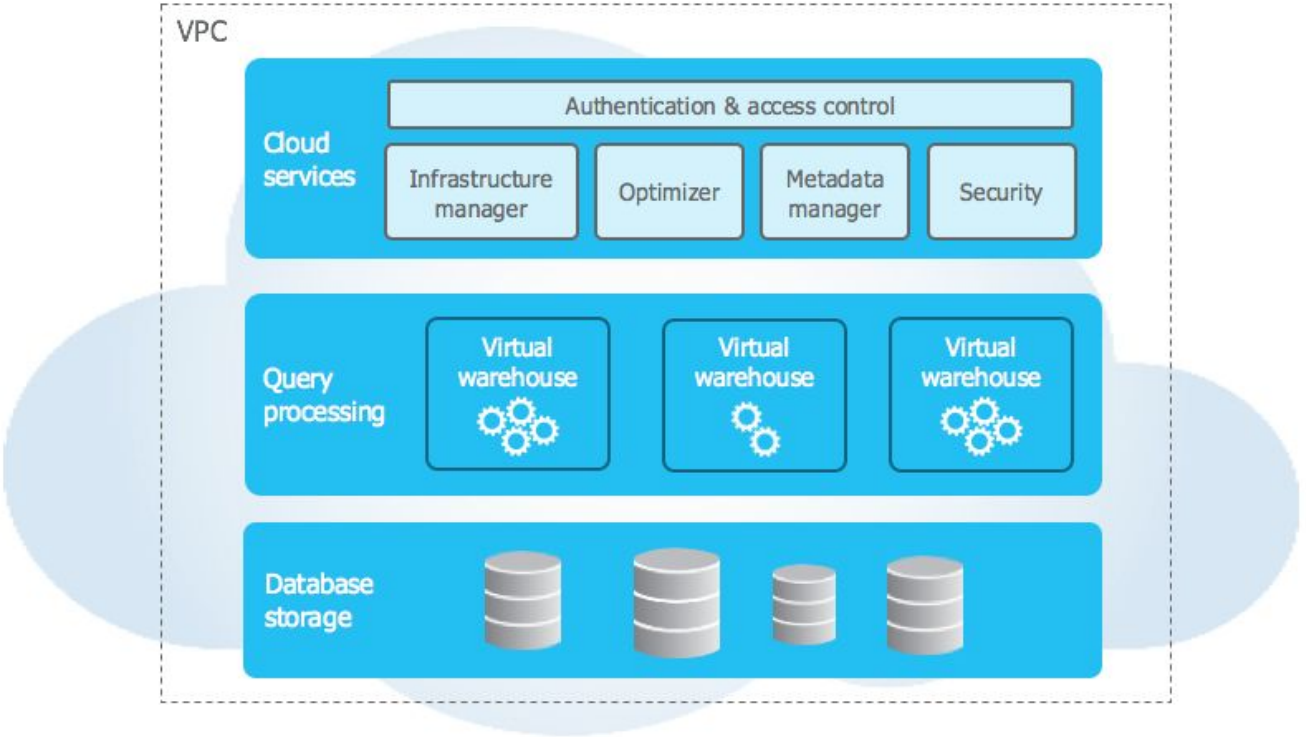
- An advanced data platform provided as a self-managed service.
- Data storage, processing, and analytic solutions that are faster, easier to use, and far more flexible than traditional offerings.
- Not built on any existing database technology.
- Combines a completely new SQL query engine with an innovative architecture natively designed for the cloud.
- Snowflake provides all of the functionality of an enterprise analytic database, along with many additional special features and unique capabilities.

Snowflake's Data Cloud: a true self-managed service

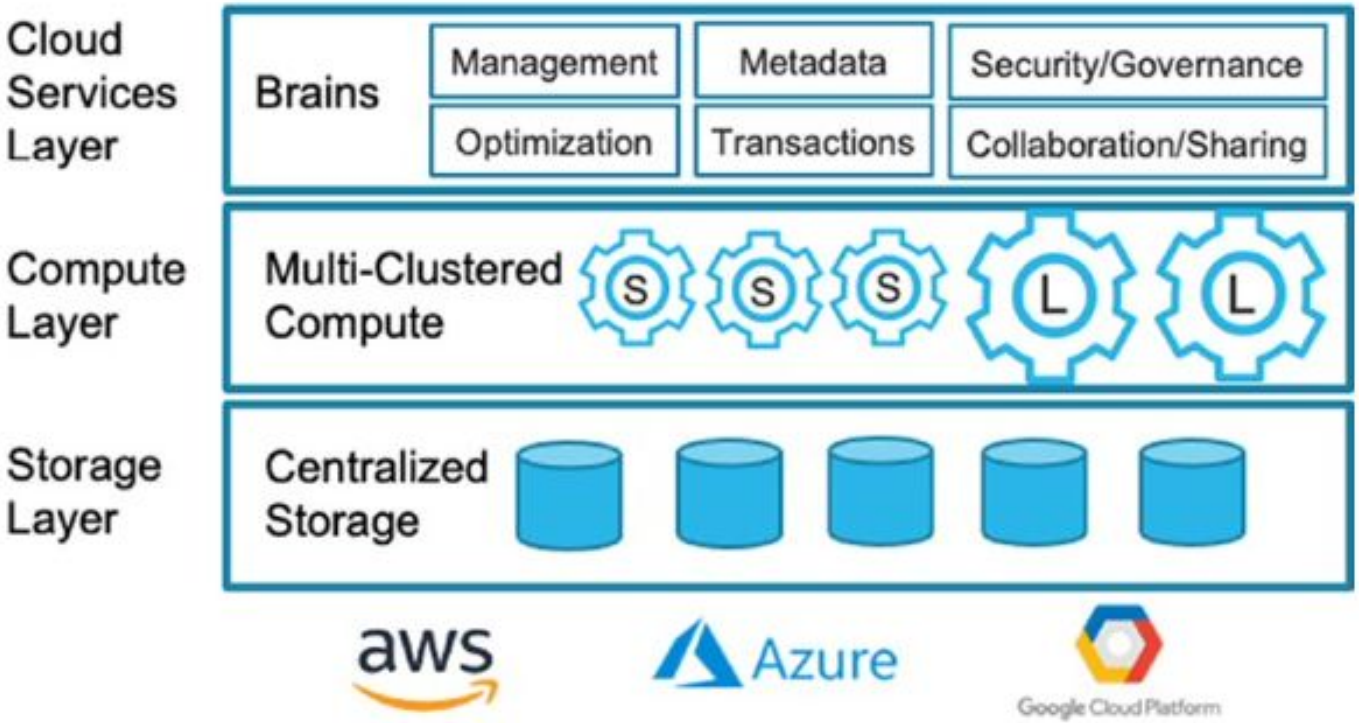
Data Platform as a Self-managed Service

- No hardware (virtual or physical) to select, install, configure, or manage.
- Virtually no software to install, configure, or manage.
- Ongoing maintenance, management, upgrades, and tuning are handled by Snowflake.

Snowflake's Data Cloud: Architecture



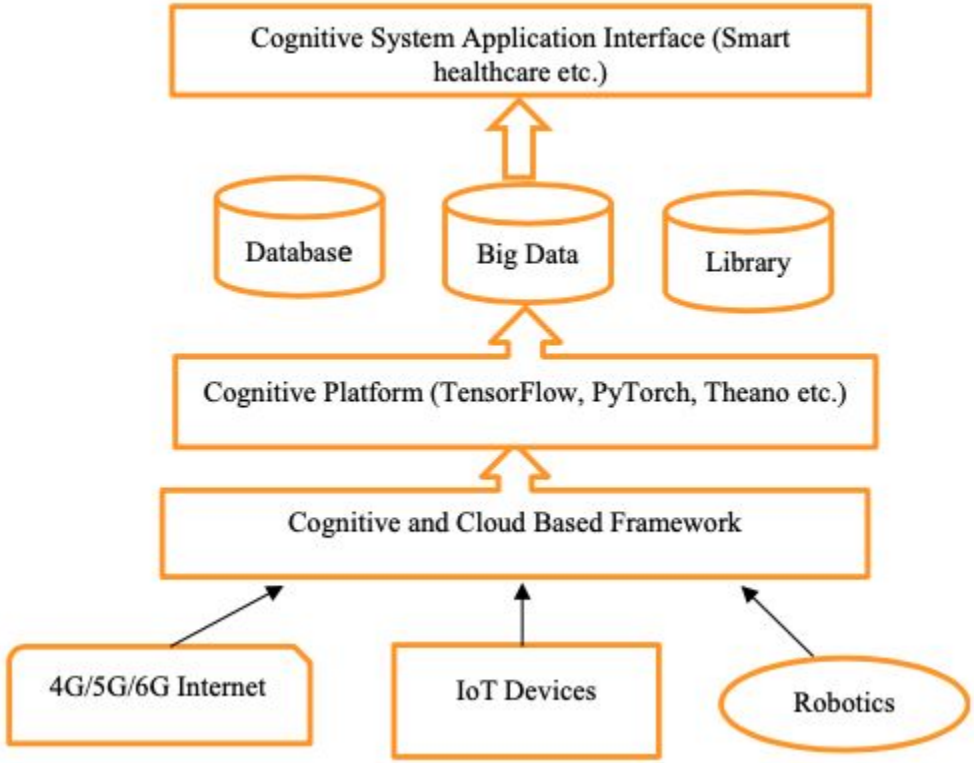
Snowflake Three Layer Architecture:



Snowflake: Unique Advantages

- The separation of compute from storage
- Automated data maintenance and scaling
- Ease of use
- Speed centric architecture
- Data sharing

Caution: No Snake Oil



Big Data: Next Steps

- Build skills for big data
  - Computational power
  - Coding efficiency
  - Statistical, Machine Learning & Deep Learning Models
- Practice (on free credits)
  - Snowflake DB
  - Data Bricks

Questions?

Connect:
revendra.iisc@gmail.com
Youtube.com/@revendrat

References

- Matter, U. (2023). Big Data Analytics: A Guide to Data Science Practitioners Making the Transition to Big Data. CRC Press.
- Databricks Documentation: https://docs.databricks.com/en/index.html
- Snowflake Documentation: https://docs.snowflake.com/en/user-guide/intro-key-concepts