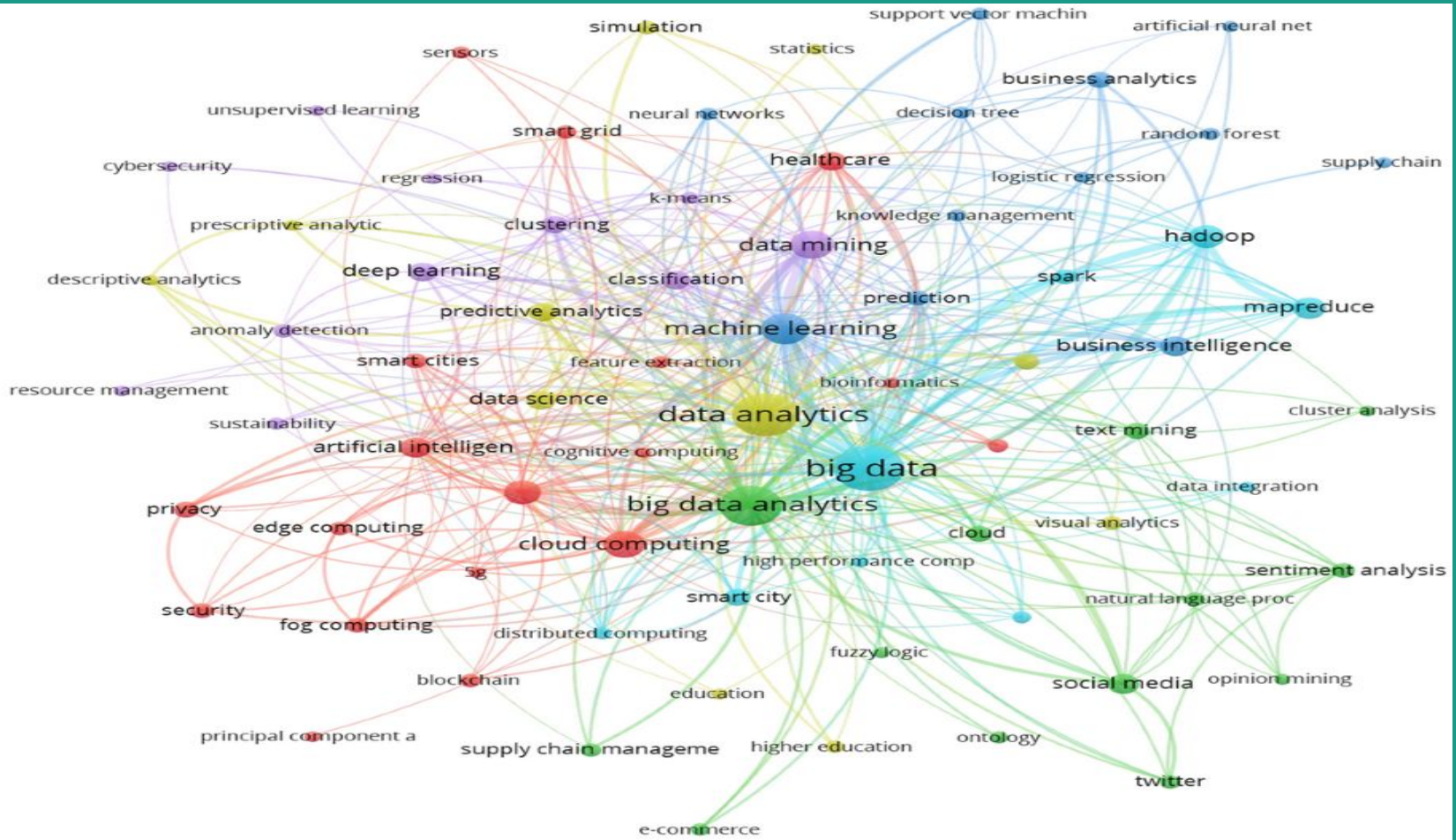

Introduction to Big Data
Recap & Conclusion

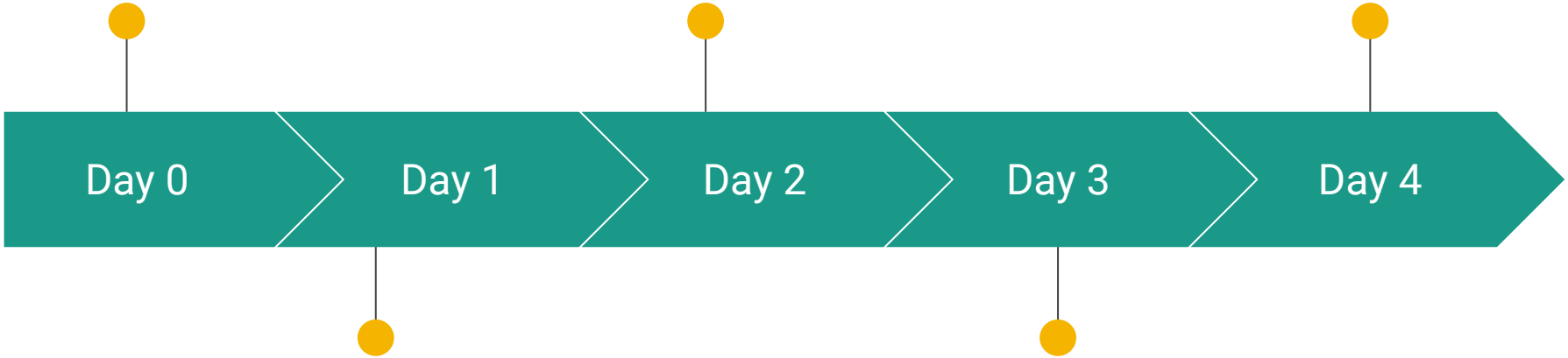

simulation
support vector machin
artificial neural net
sensors
statistics
business analytics
unsupervised learning
neural networks
decision tree
smart grid
random forest
cybersecurity
healthcare
supply chain
regression
logistic regression
k-means
knowledge management
prescriptive analytic
clustering
hadoop
data mining
deep learning
classification
spark
descriptive analytics
prediction
predictive analytics
mapreduce
anomaly detection
machine learning
business intelligence
smart cities
feature extraction
bioinformatics
resource management
data science
cluster analysis
data analytics
sustainability
text mining
artificial intelligen
cognitive computing
big data
data integration
privacy
big data analytics
edge computing
visual analytics
cloud
cloud computing
high performance comp
sentiment analysis
5g
smart city
natural language proc
security
fog computing
distributed computing
fuzzy logic
blockchain
social media
opinion mining
education
principal component a
supply chain manageme
higher education
ontology
twitter
e-commerce

Our Learning Journey – A Quick Review

Intro to Data Science & Big Data

Pig & Hive

...



Hadoop; HDFS & MapReduce

NoSQL & Spark

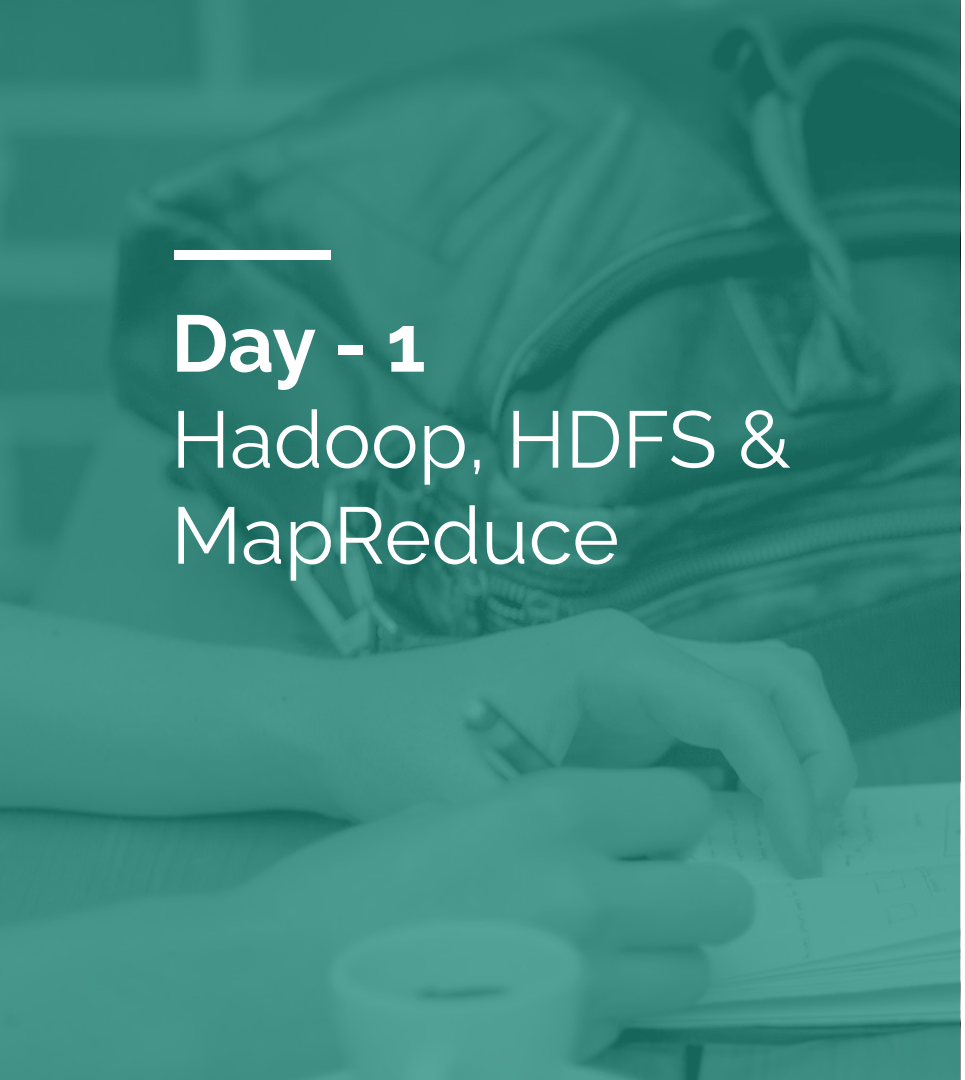
Day - 0
Intro to Data Science & Big Data

The Foundations: Why Does Big Data Matter?

Day - 0
Intro to Data Science & Big Data

The Foundations: Why Does Big Data Matter?

- The explosion of data volumes and variety
- Challenges of traditional data processing
- How Big Data technologies enable new insights and solutions



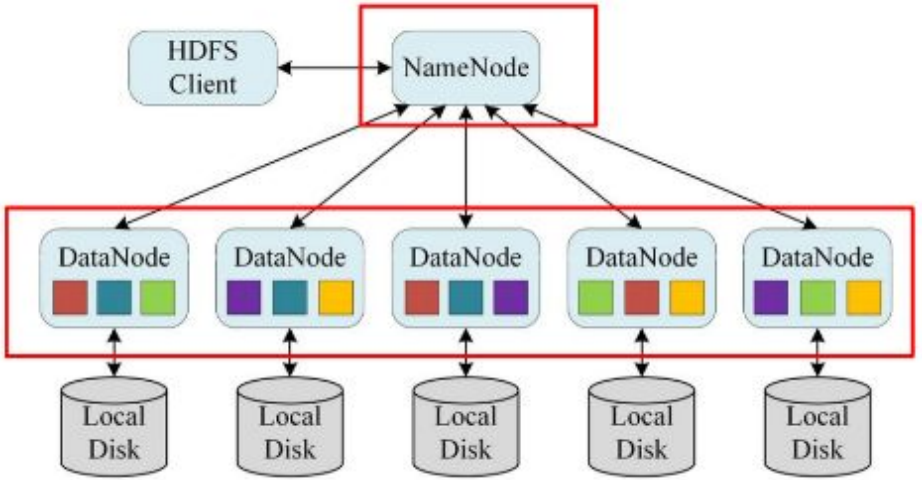
Day - 1
Hadoop, HDFS & MapReduce

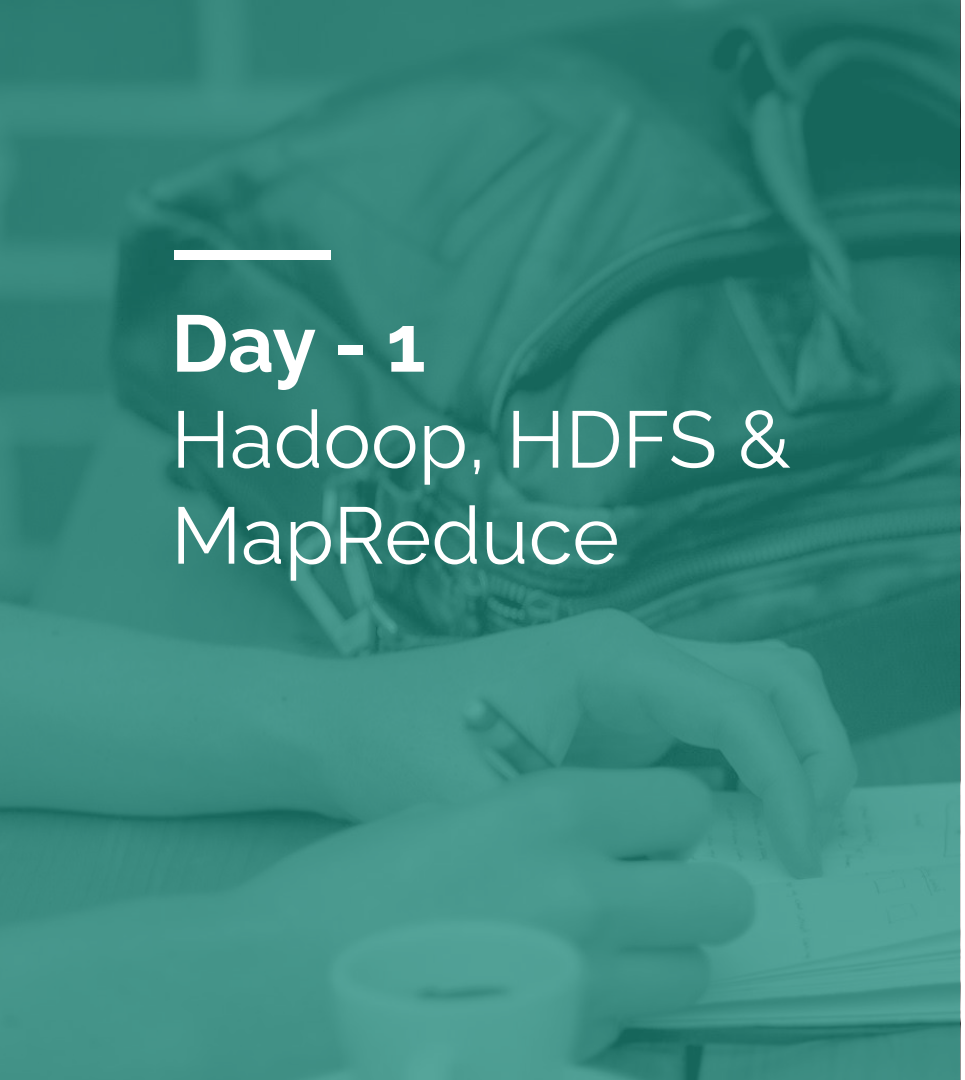
HDFS: The Cornerstone of Big Data Storage

Day - 1

Hadoop, HDFS & MapReduce

HDFS: The Cornerstone of Big Data Storage

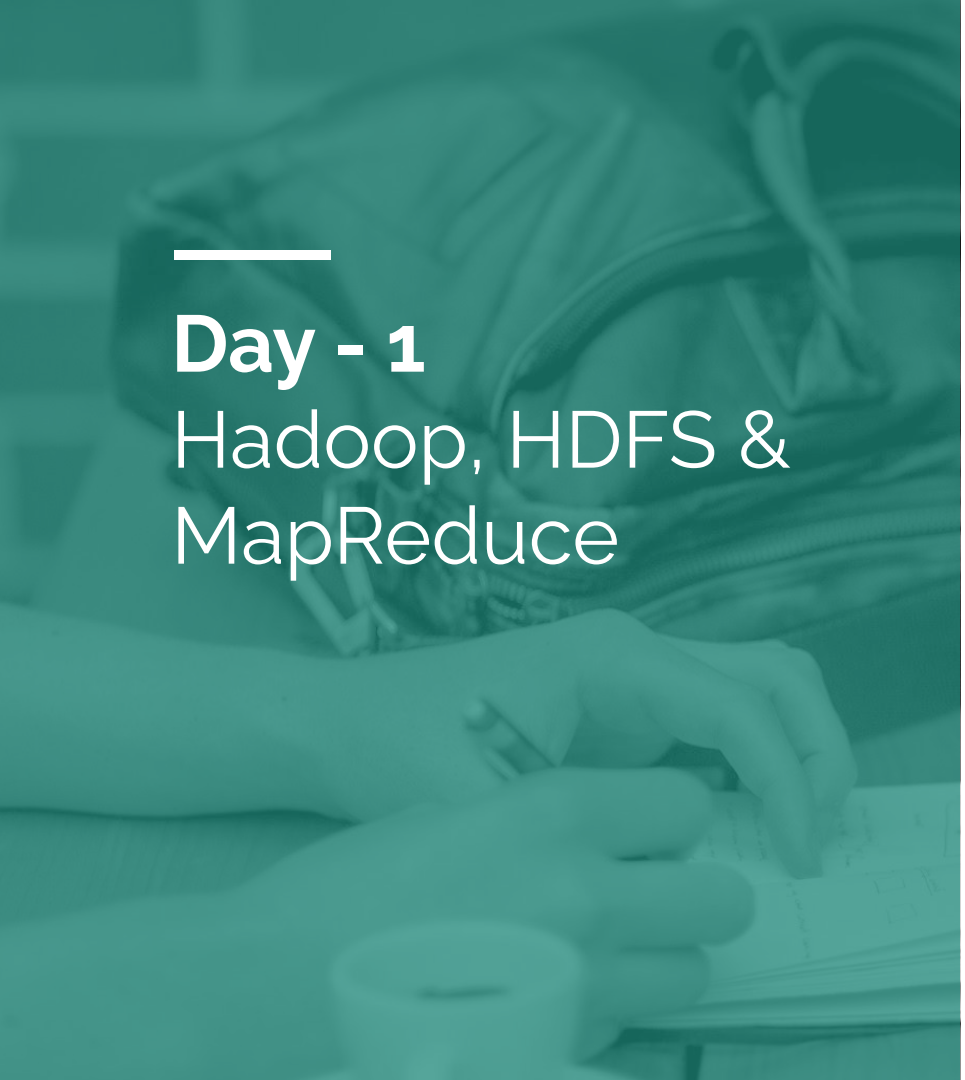




Day - 1
Hadoop, HDFS & MapReduce

HDFS: The Cornerstone of Big Data Storage

- HDFS Architecture (NameNode, DataNodes, Replication)
- Scalability and fault tolerance provided by HDFS
- Key concepts: blocks, distributed storage



Day - 1

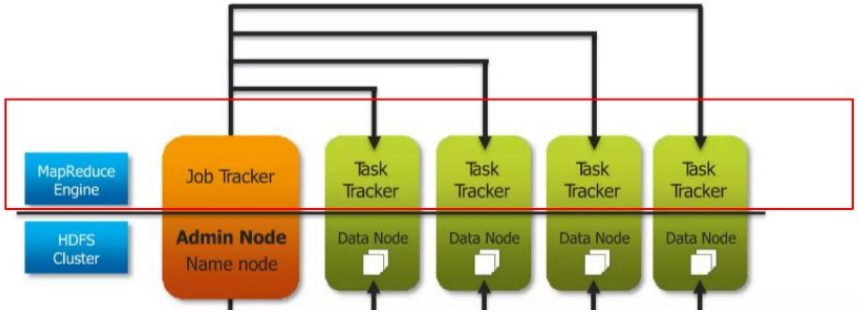
Hadoop, HDFS & MapReduce

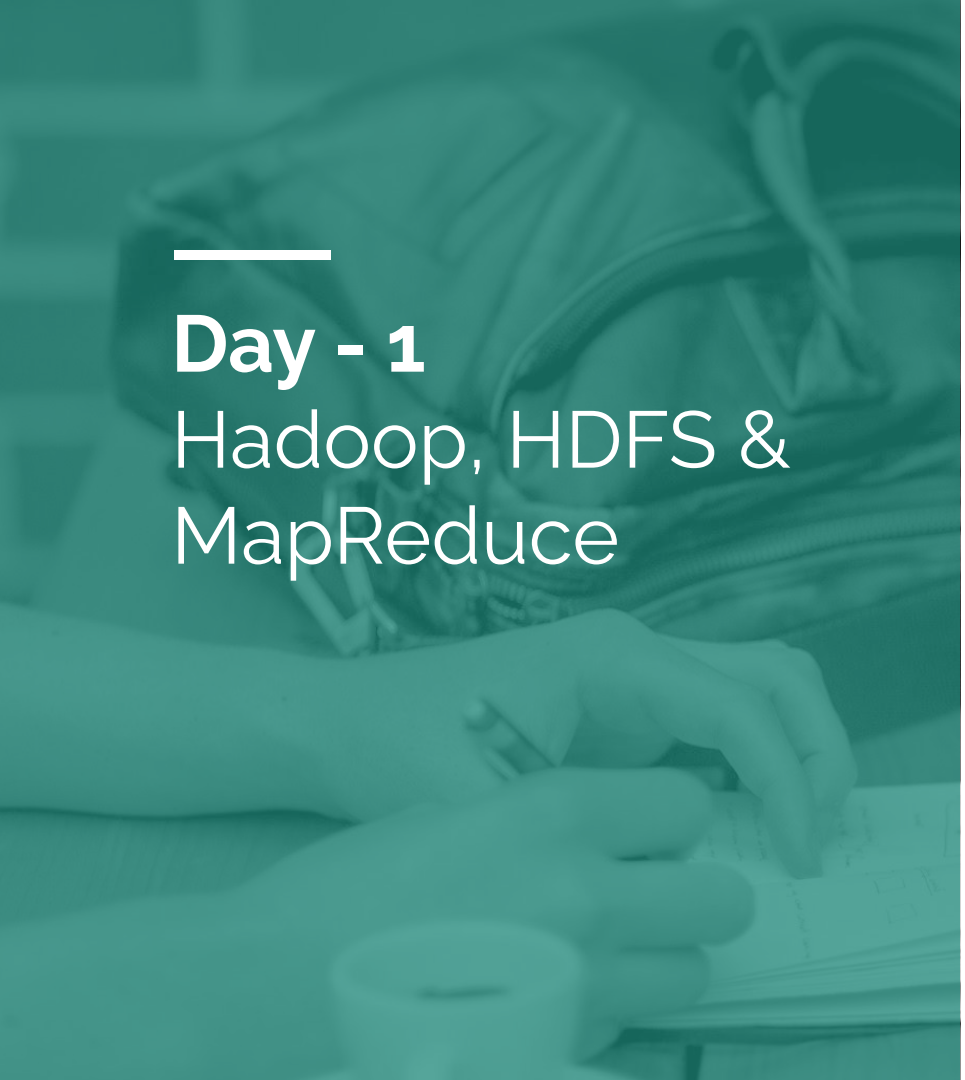
MapReduce: The Engine of Distributed Processing

Day - 1

Hadoop, HDFS & MapReduce

MapReduce: The Engine of Distributed Processing

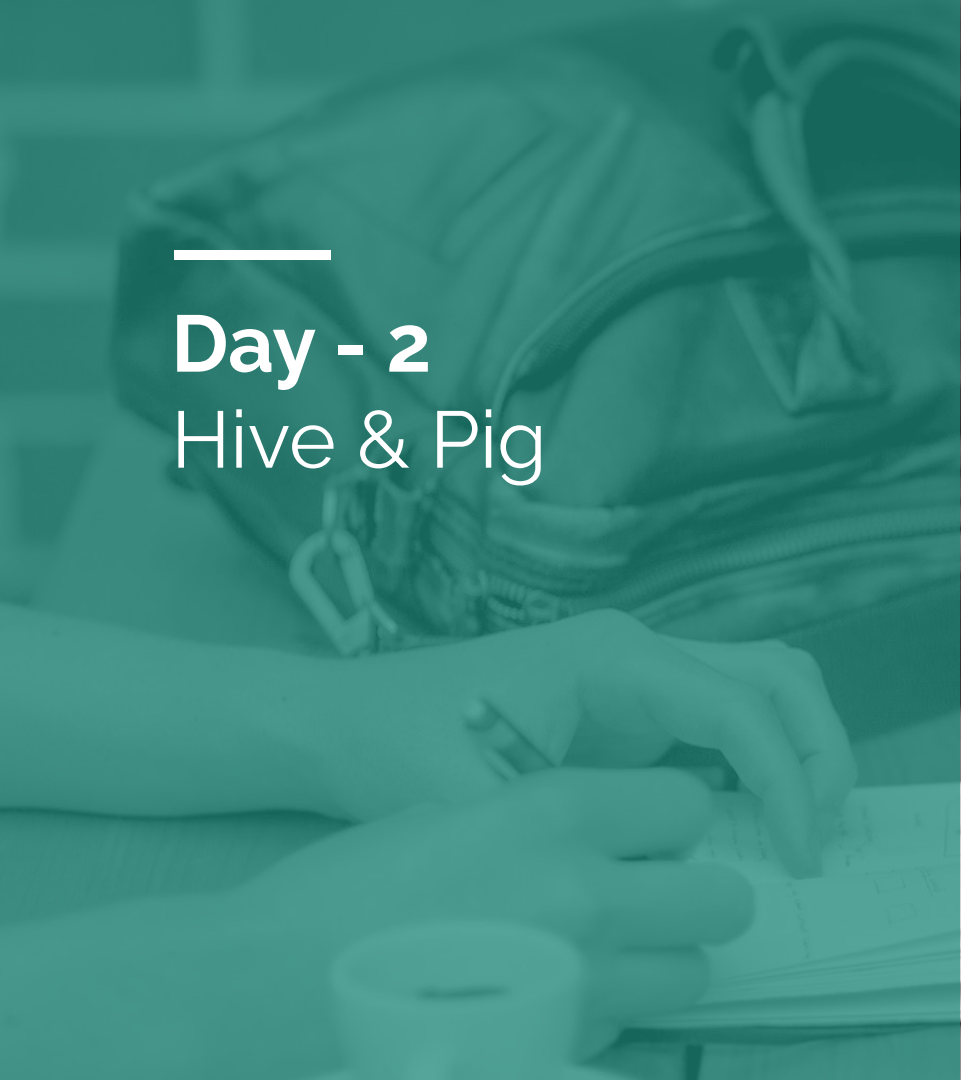




Day - 1
Hadoop, HDFS & MapReduce

MapReduce: The Engine of Distributed Processing

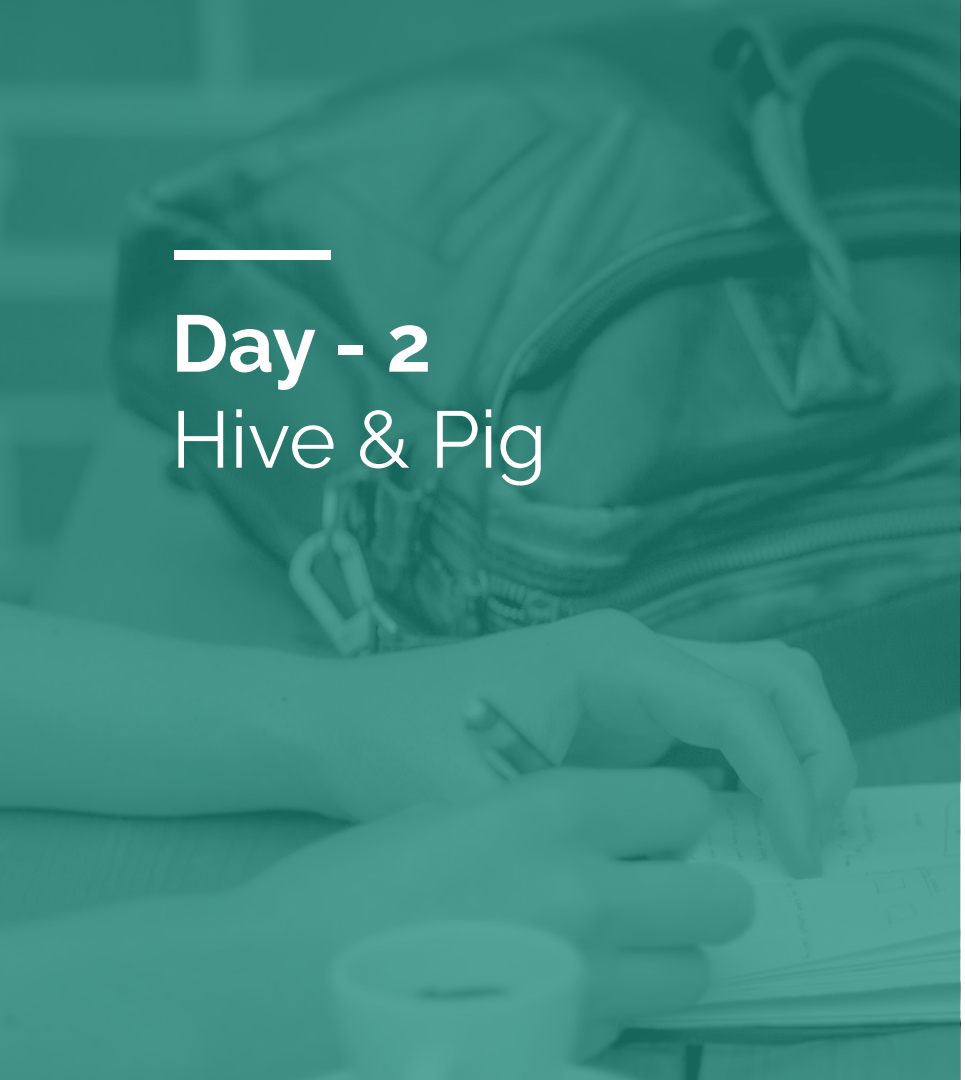
- Key Idea: Breaking down large-scale computations into smaller units.
- Core Stages: Map Phase, Shuffle & Sort, Reduce Phase



Day - 2

Hive & Pig

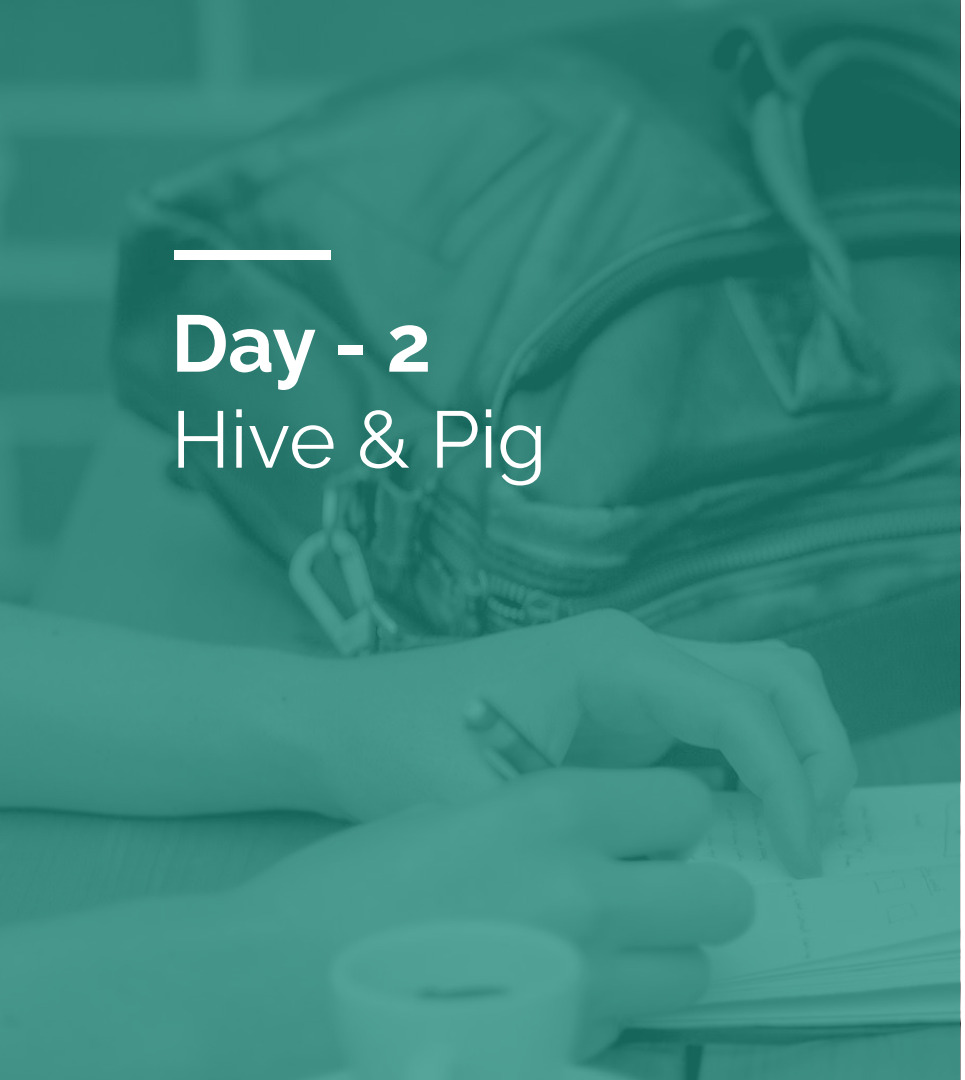
Hive: Bringing SQL-like Structure to Hadoop



Day - 2
Hive & Pig

Hive: Bringing SQL-like Structure to Hadoop

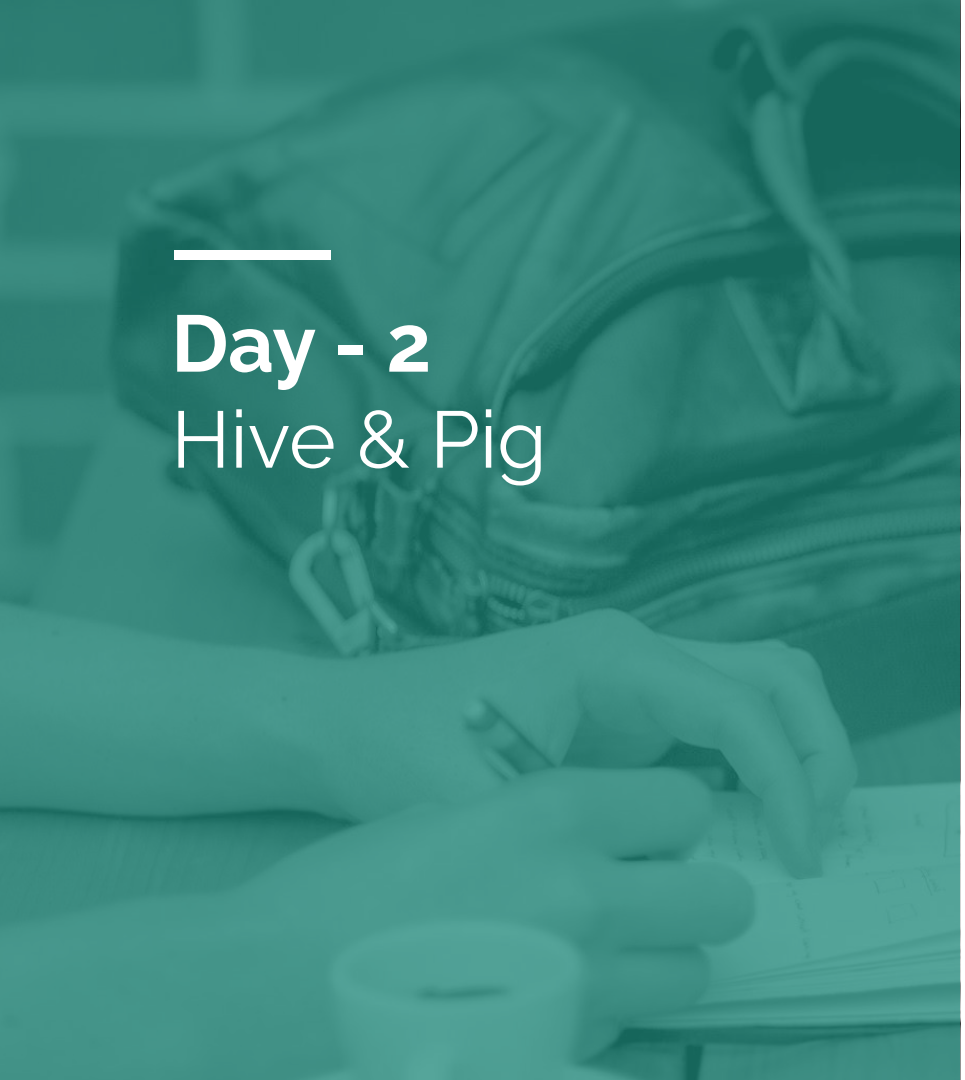
- Provides familiar SQL interface for querying data in HDFS
- Uses abstractions over MapReduce (you don't write MapReduce code directly)
- Ideal for analysts comfortable with SQL



Day - 2
Hive & Pig

Pig: Data Flow for Power Users

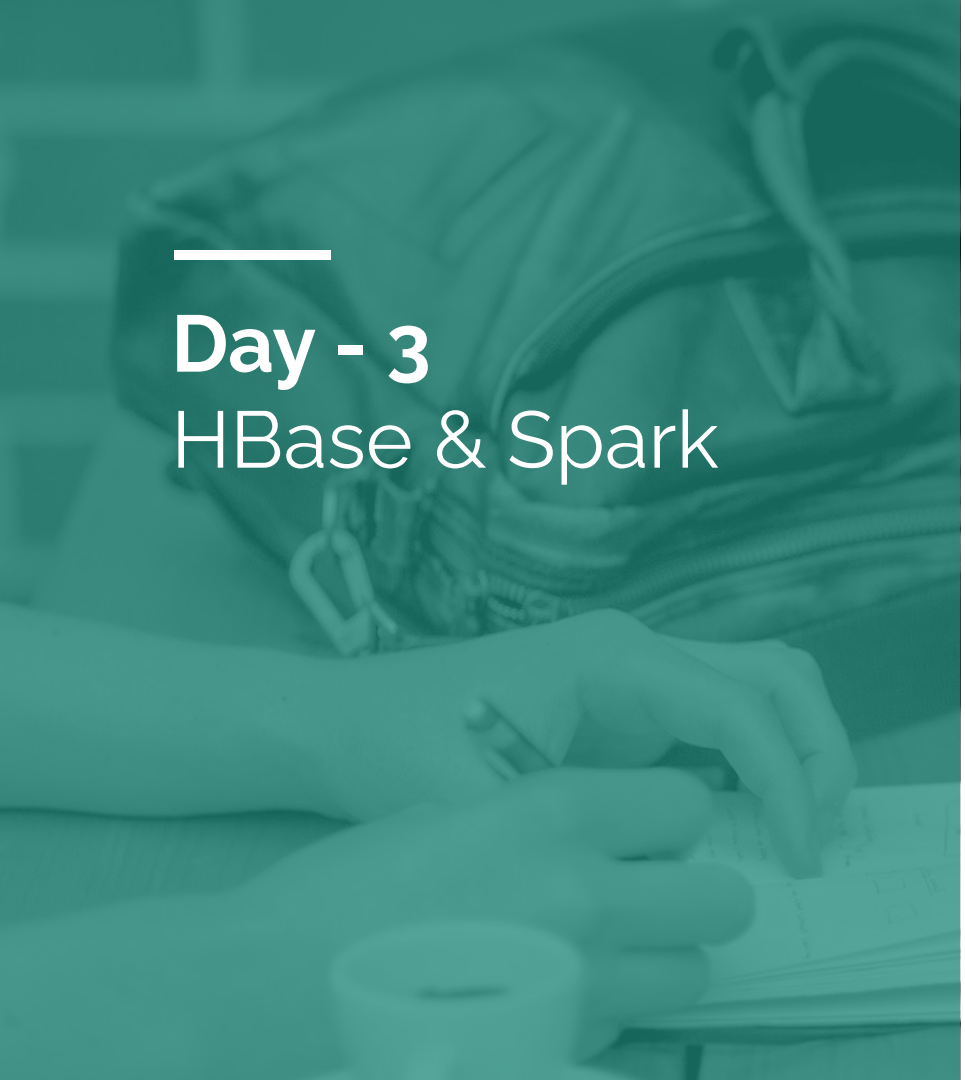
Apache Pig



Day - 2
Hive & Pig

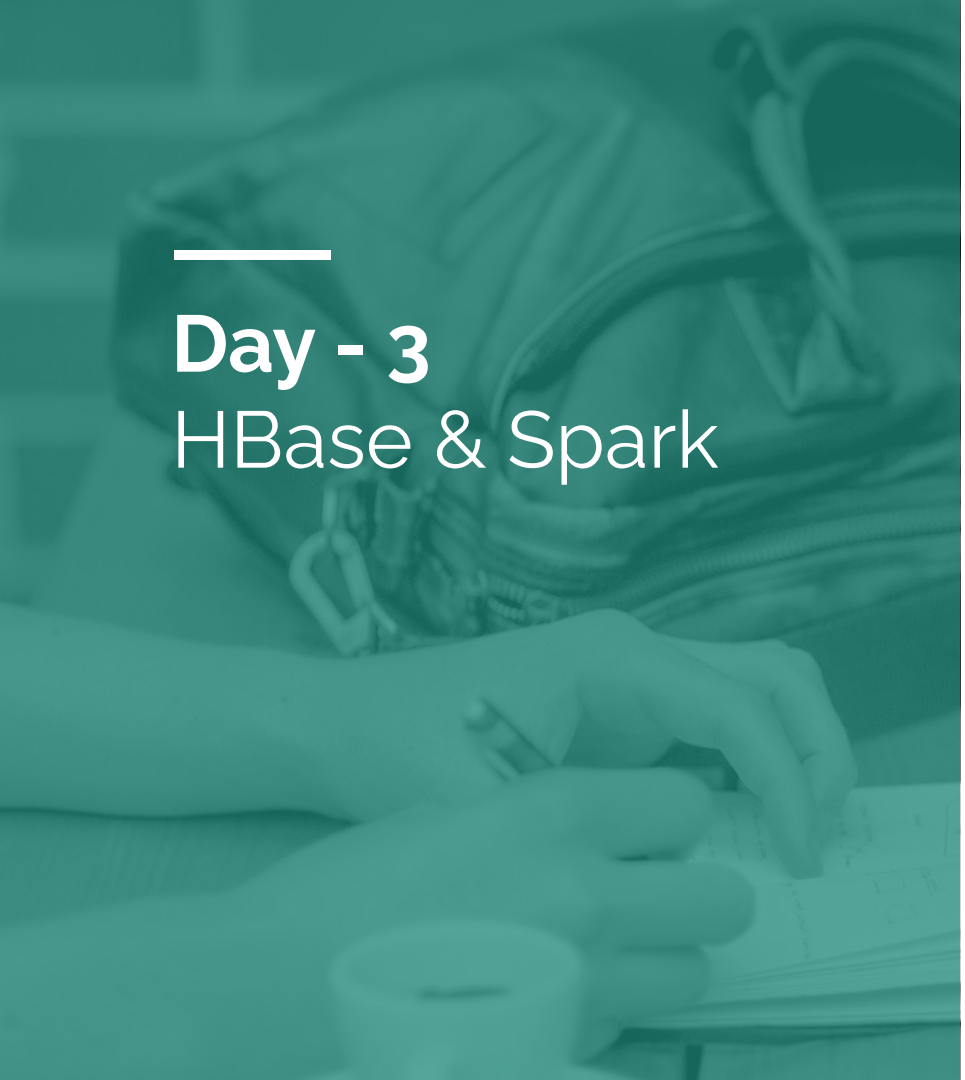
Pig: Data Flow for Power Users

- Pig Latin: A procedural language for transforming data
- Focus on chaining operations (Load, Filter, Join, etc.) to build pipelines
- Well-suited for complex data cleaning and preparation tasks



Day - 3
HBase & Spark

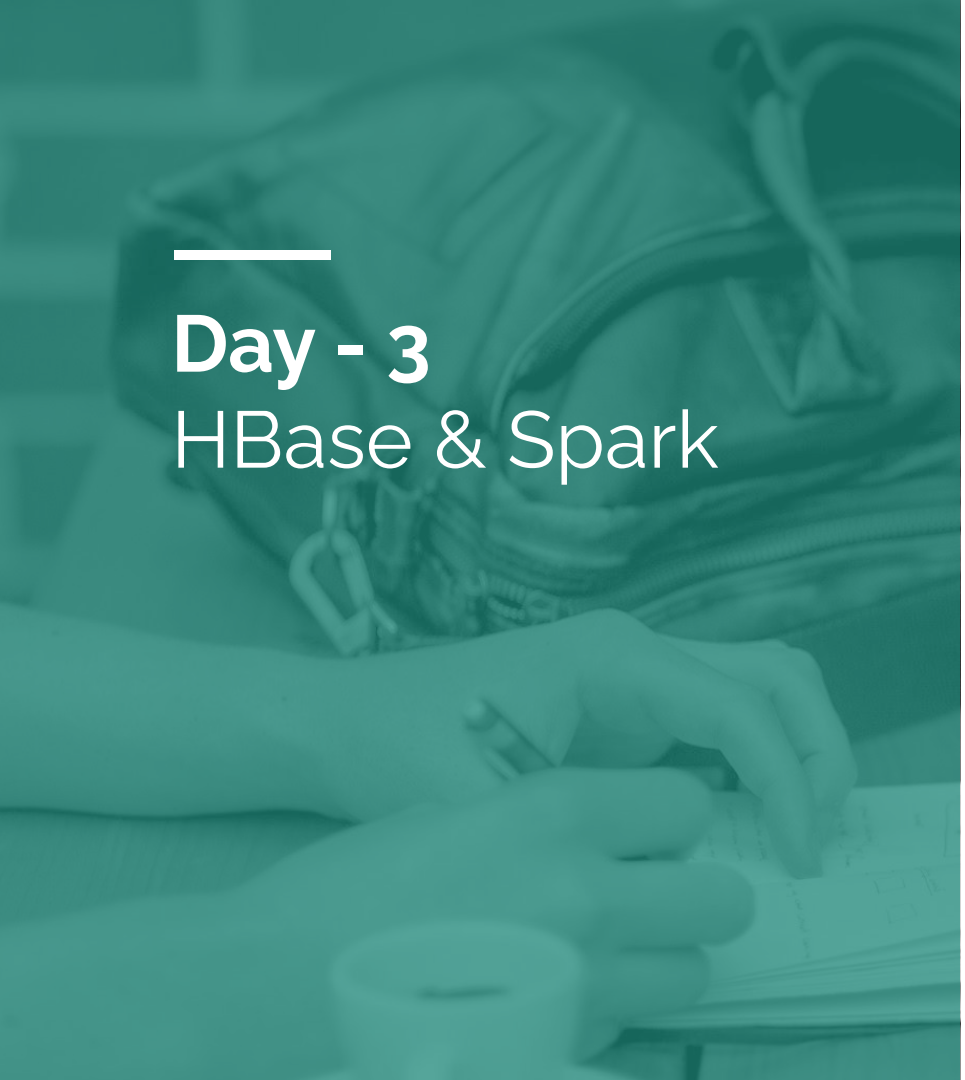
HBase: Real-time, Scalable NoSQL Database



Day - 3
HBase & Spark

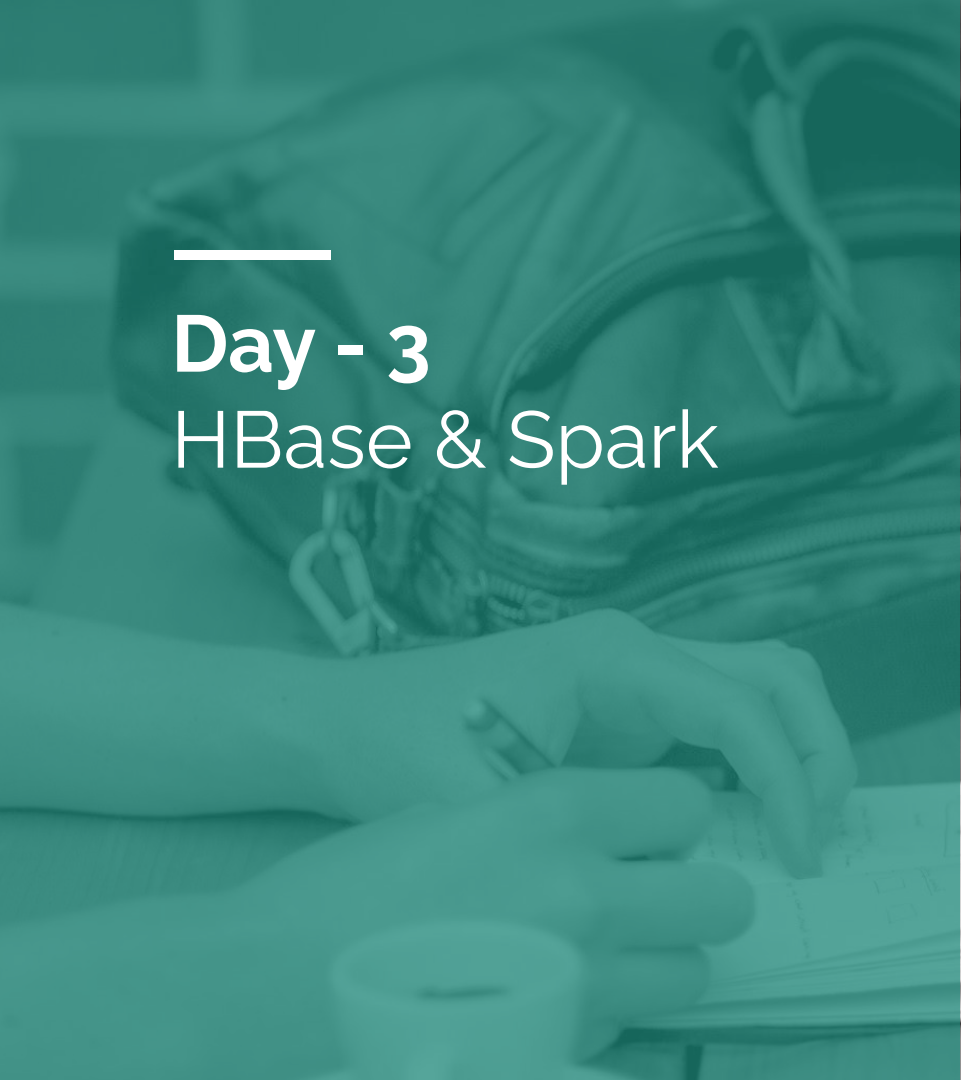
HBase: Real-time, Scalable NoSQL Database

- Open-source, column-oriented database built on top of Hadoop (HDFS).
- Modeled after Google's Bigtable.
- Supports sparse data storage.



Day - 3
HBase & Spark

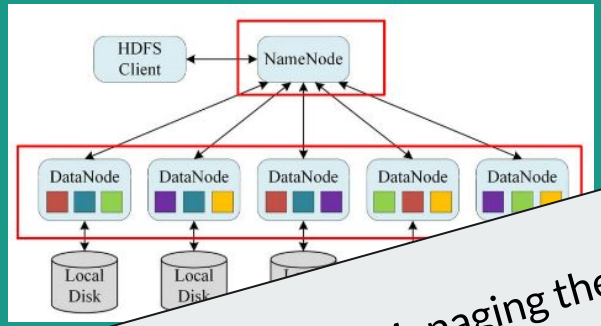
Spark: Fast, Flexible, and Beyond MapReduce

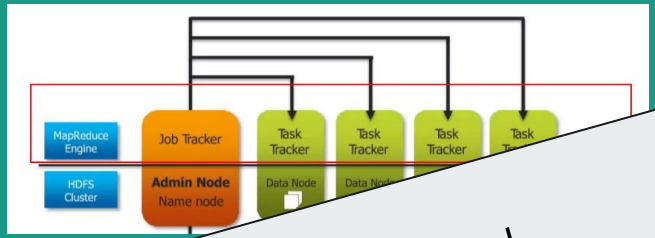


Day - 3
HBase & Spark

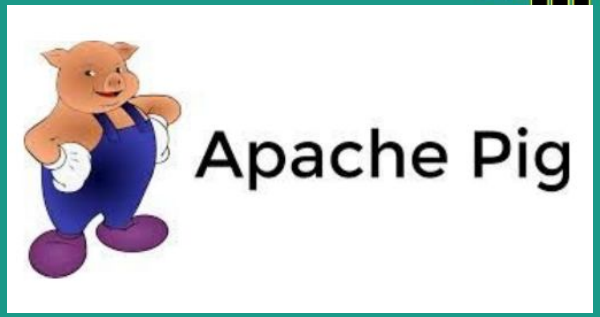
Spark: Fast, Flexible, and Beyond MapReduce

- In-memory processing for speed
- Expanded capabilities: Streaming, machine learning, graph analysis
- RDDs (Resilient Distributed Datasets) as the primary data abstraction





Managing them all individually can be complex!



Big Data & Cloud Computing: A Perfect Match!

Harnessing the Cloud for Big Data

Scalability, Agility, and Cost-Efficiency

Why Cloud Computing for Big Data?

The Cloud Advantage

Scalability: Easily add/remove resources to match data volume and processing needs.

Cost-effectiveness: Pay-as-you-go models eliminate upfront infrastructure investments.

Agility: Quicker to deploy and experiment with big data tools and services.

Innovation: Access the latest in AI/ML technologies often offered as cloud services.

Cloud Big Data Services: The Building Blocks

Your Cloud Big Data Toolkit

- **AWS - Amazon**
- **Azure - Microsoft**
- **GCP - Google**

Cloud Big Data Services: The Building Blocks

Your Cloud Big Data Toolkit

- **Storage:**
  - **S3 (AWS)**
  - **Azure Blob Storage**
  - **Google Cloud Storage**
- **Compute:**
  - **EC2 (AWS)**
  - **Azure Virtual Machines**
  - **Google Compute Engine**

Cloud Big Data Services: The Building Blocks

Your Cloud Big Data Toolkit

- **Processing Frameworks:**
  - **EMR (AWS)**
  - **HDInsight (Azure)**
  - **Dataproc (GCP)**
- **Analytics & ML:**
  - **BigQuery (GCP)**
  - **Redshift (AWS)**
  - **Azure Synapse**

Demo - GCP

Uploading Data to Google Cloud Storage (GCS)

Google Cloud Platform

- Navigate to the GCS console: Find the GCS in the navigation menu.
- Create a Bucket: Create a new bucket with a suitable name and storage settings.
- Upload Dataset: Upload the sales data.

Data Exploration & Analysis with BigQuery

Google Cloud Platform

- Navigate to the BigQuery console
- Create a Dataset: Create a new dataset to hold our query results.
- Load from Storage: Run a query to load the sales data directly from GCS into a BigQuery table.
- Write SQL Queries
- Visualizing Results: Use BigQuery's built-in charts.

Data Exploration & Analysis with BigQuery

Google Cloud Platform

Create and **train models** directly within BigQuery using SQL-like syntax.

```
CREATE OR REPLACE MODEL
sales_analysis.sales_forecast_model

OPTIONS (model_type='linear_reg') AS

SELECT

  timestamp,

  region,

  sales_amount

FROM sales_analysis.sales_data;
```

Processing with Dataproc

Google Cloud Platform

Creating a Cluster: Set up a simple Dataproc cluster, configuring the type and number of machines.

https://medium.com/google-cloud/all-you-need-to-know-about-google-cloud-dataproc-23fe91369678

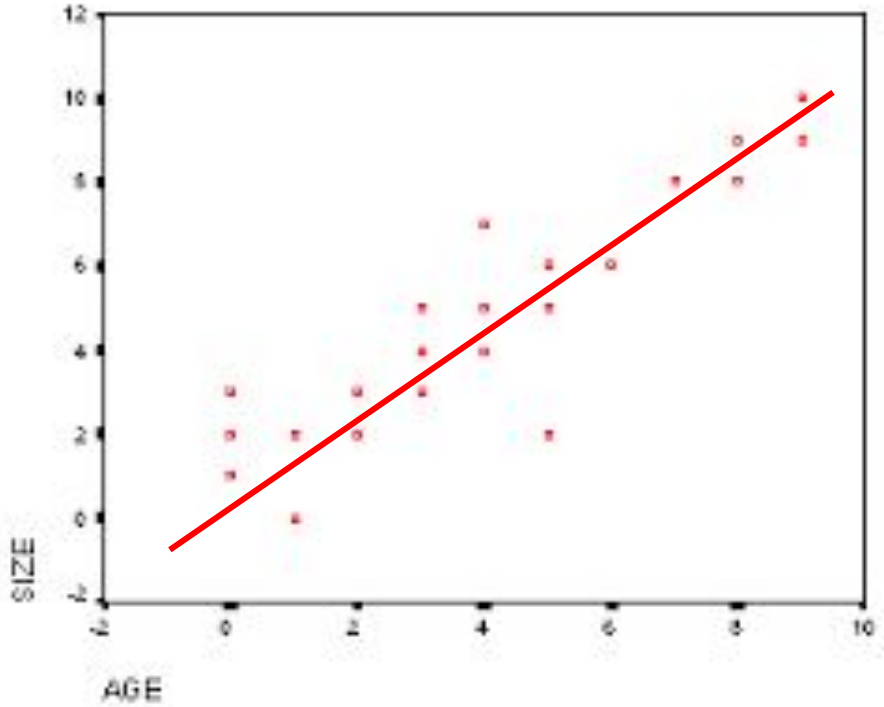
Understanding Linear Regression

What is Linear Regression?

- A simple technique for finding relationships in data.
- Helps us predict one thing based on another.
- Think of drawing the best-fit line through a bunch of data points.

What is Linear Regression?

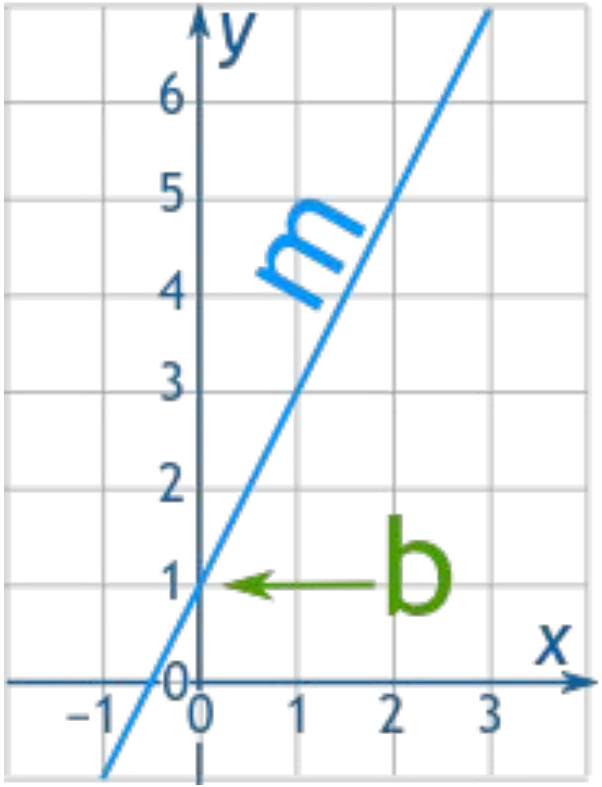
Think of drawing the best-fit line through a bunch of data points.

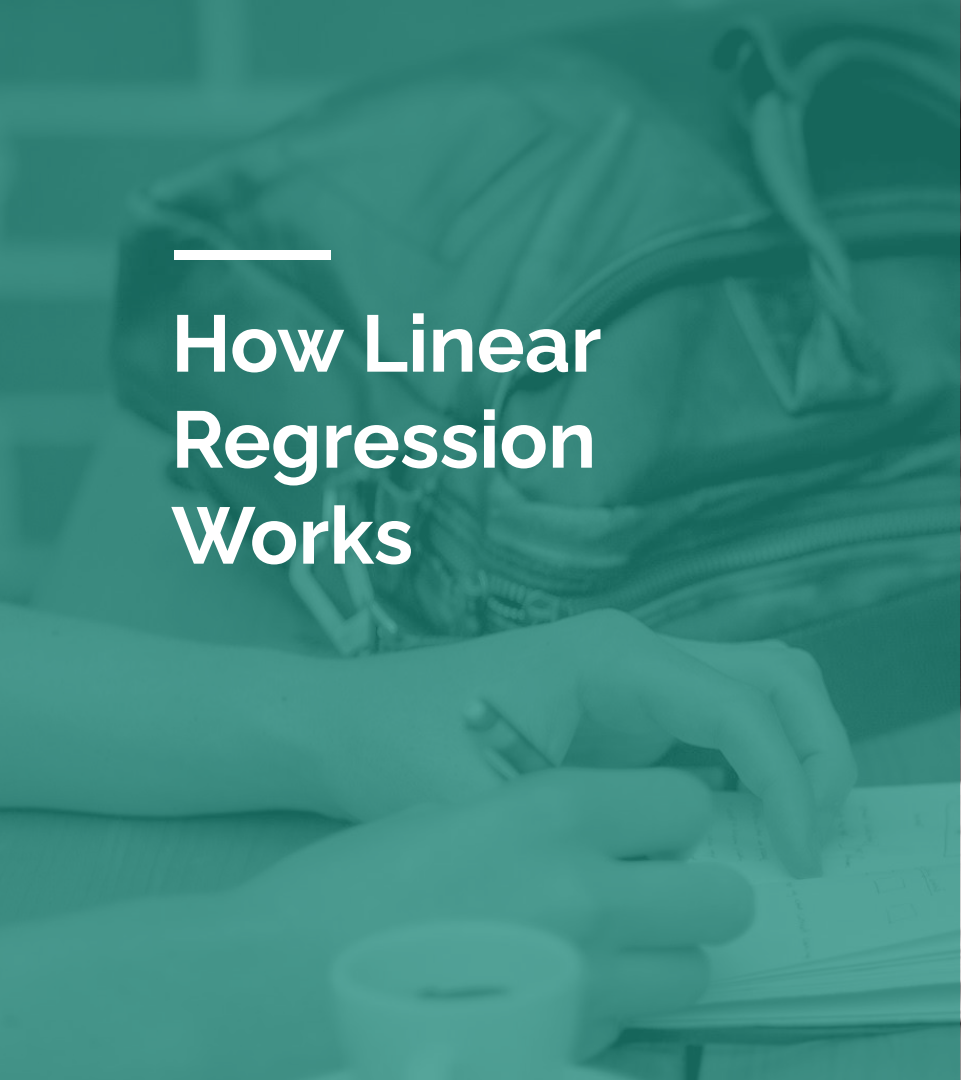


How Linear Regression Works

$$y = mx + b$$

The Magic Equation

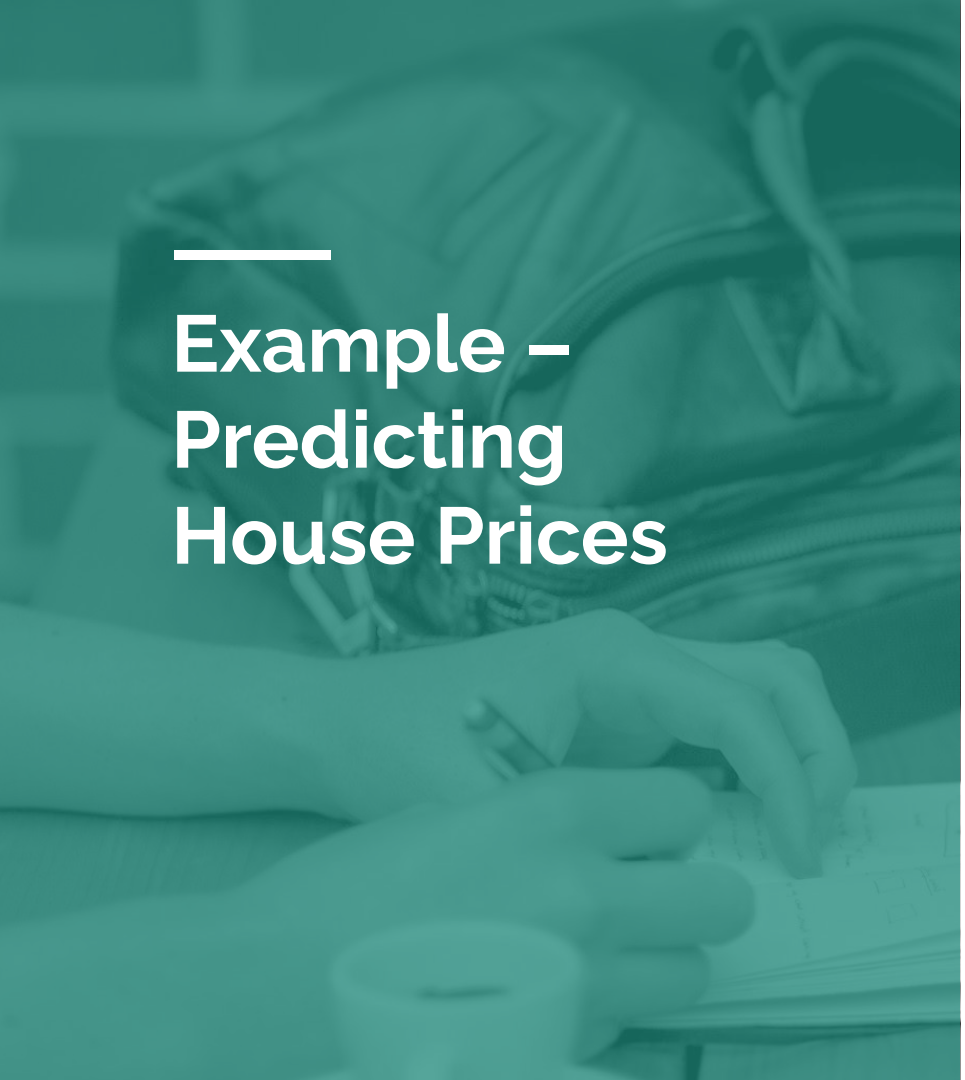




How Linear Regression Works

The Magic Equation

- The line is represented by the equation: y = mx + b
- 'y' is the thing we want to predict (like sales)
- 'x' is the thing we use to predict (like advertising spending)
- 'm' is the slope of the line (how steep it is)
- 'b' is where the line crosses the y-axis (the starting point)



Example – Predicting House Prices

The Magic Equation

- **Data**: Square footage of houses and their selling prices.
- Linear regression finds a relationship between size and price.
- The line can help estimate the price of a house if we know its size.

Quiz

https://forms.gle/BpMiZWA99mz5DhL36

Thank You!