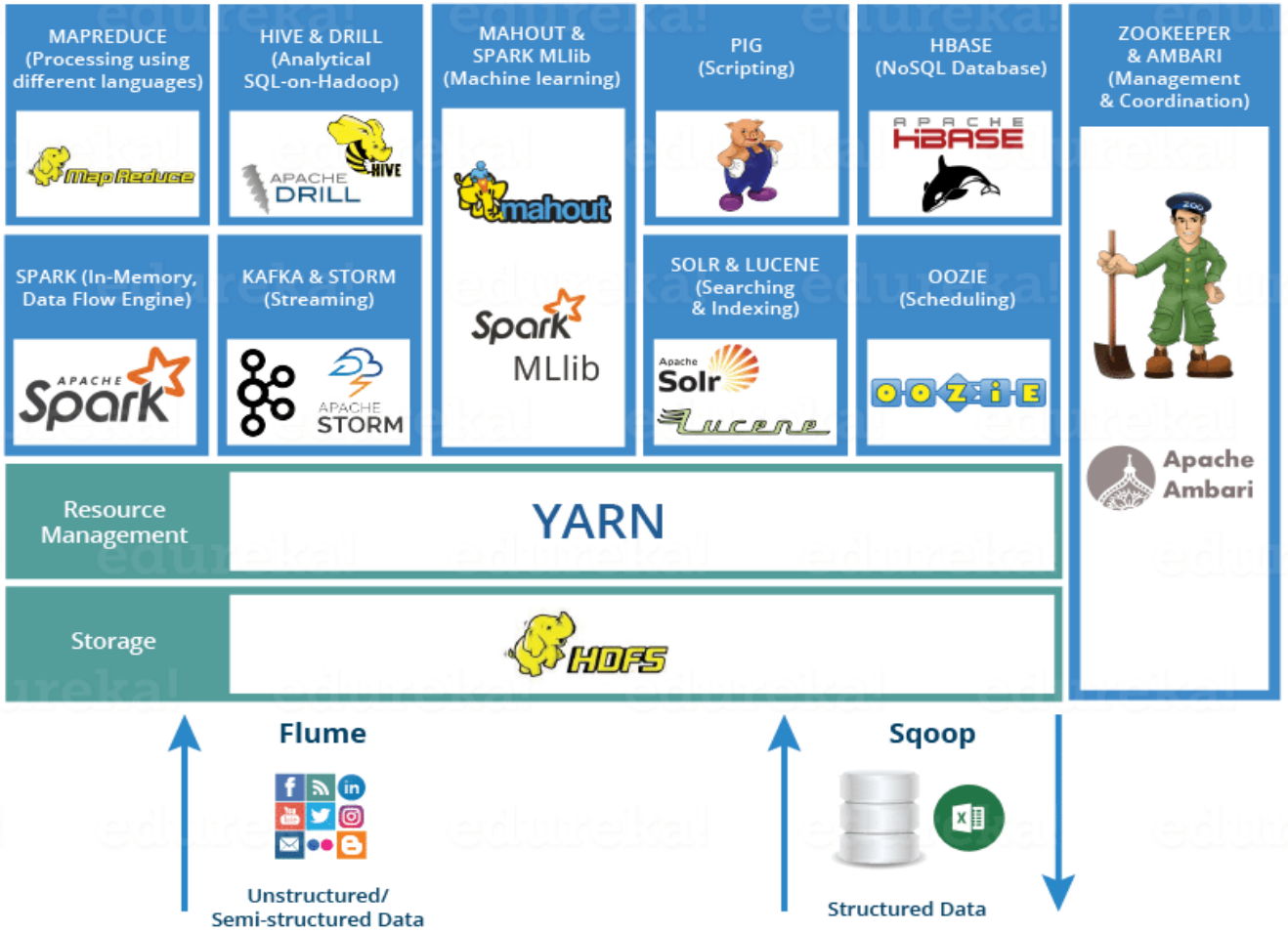
Introduction to Basics of Big Data

Hadoop Ecosystem



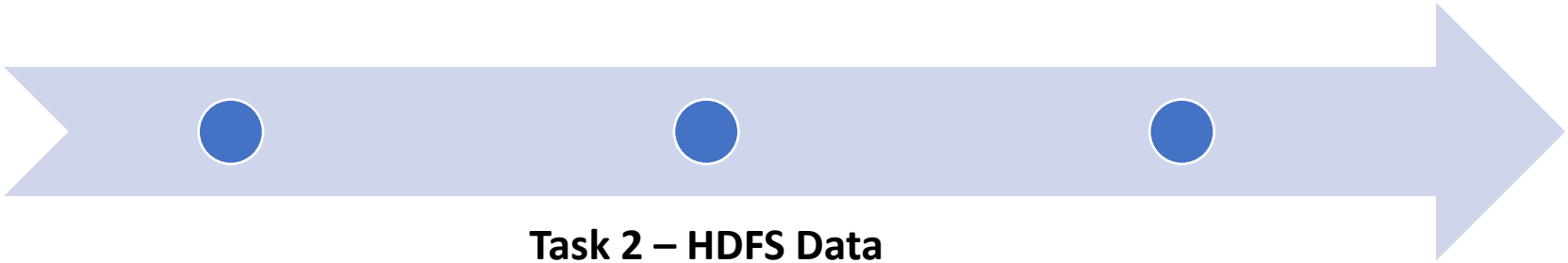
HDFS -> Hadoop Distributed File System
YARN -> Yet Another Resource Negotiator
MapReduce -> Data processing using programming
Spark -> In-memory Data Processing
PIG, HIVE-> Data Processing Services using Query (SQL-like)
HBase -> NoSQL Database

Mahout, Spark MLlib -> Machine Learning
Apache Drill -> SQL on Hadoop
Zookeeper -> Managing Cluster
Oozie -> Job Scheduling
Flume, Sqoop -> Data Ingesting Services
Solr & Lucene -> Searching & Indexing
Ambari -> Provision, Monitor and Maintain cluster

Learning Roadmap

Task 1 – Introduction to Hadoop & HDFS

Objective: Understand Hadoop architecture, HDFS concepts, block storage, and replication.



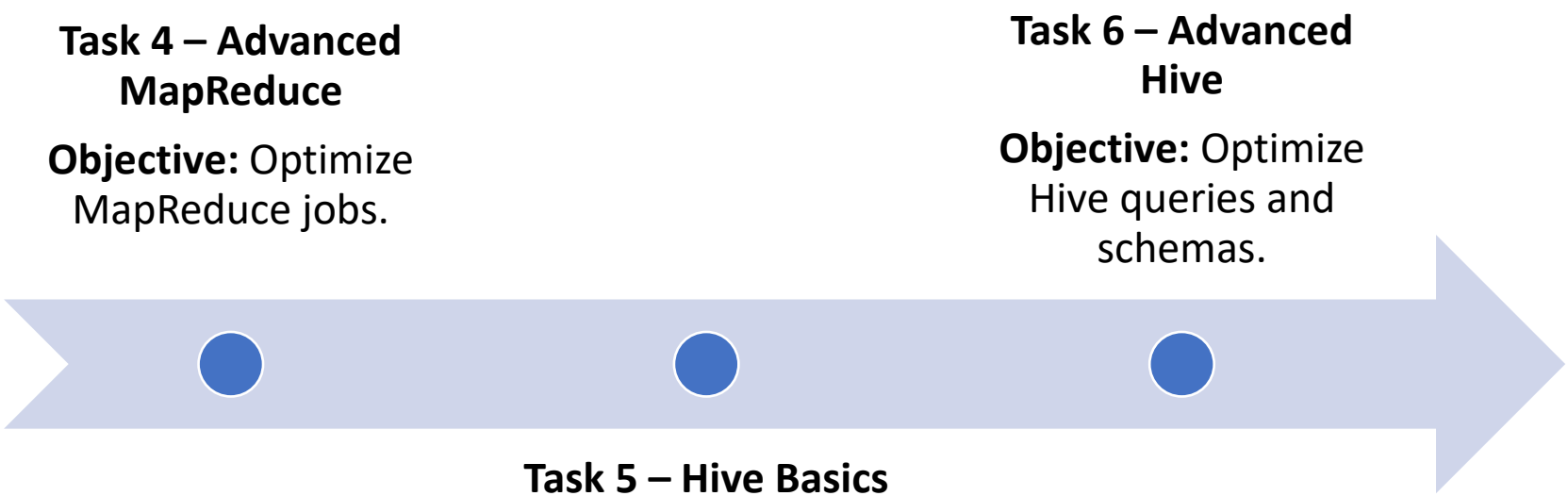
Task 2 – HDFS Data Management

Objective: Manage large-scale datasets in HDFS.

Task 3 – MapReduce Basics

Objective: Learn distributed computation with MapReduce.

Learning Roadmap



Learning Roadmap

Task 7 – Apache Pig

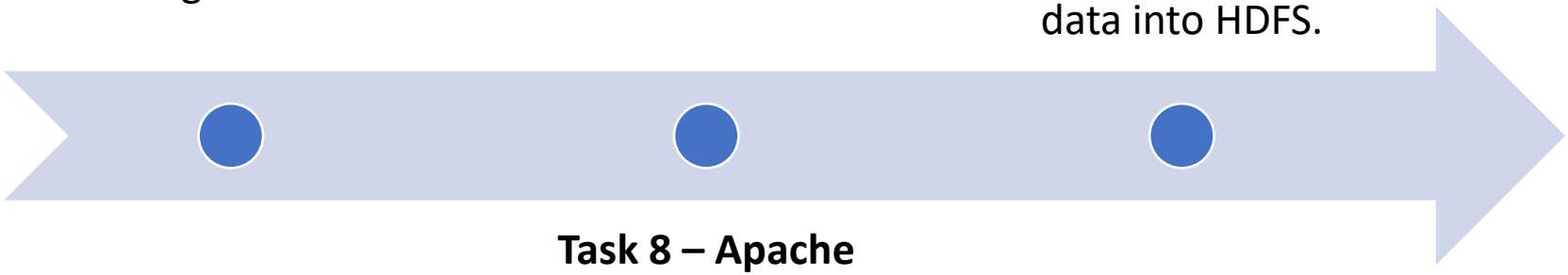
Objective: Perform ETL operations using Pig Latin.

Task 8 – Apache Sqoop

Objective: Transfer data between RDBMS and Hadoop.

Task 9 – Apache Flume

Objective: Ingest log data into HDFS.



Learning Roadmap

Task 10 – Zookeeper Basics

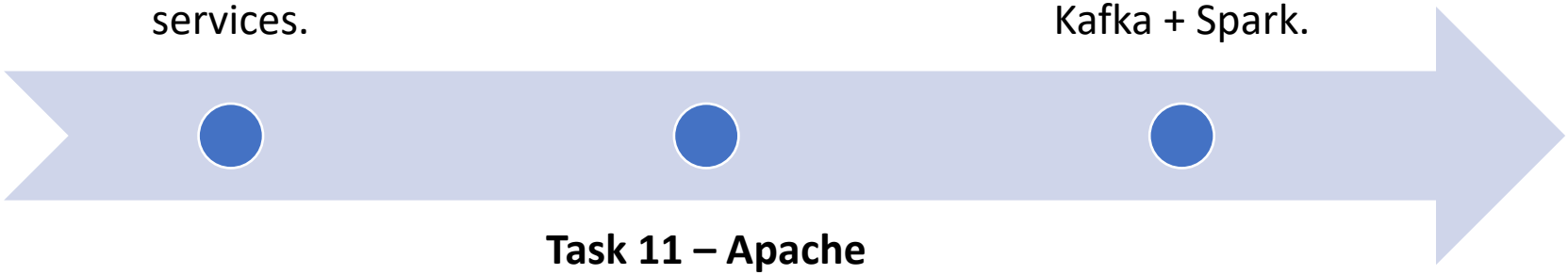
Objective: Understand Zookeeper's role in coordination services.

Task 11 – Apache Kafka Basics

Objective: Learn distributed messaging for streaming pipelines.

Task 12 – Kafka with Spark Streaming

Objective: Real-time data processing using Kafka + Spark.



Learning Roadmap

Task 13 – ELK Stack (Elasticsearch, Logstash, Kibana)

Objective: Build a real-time log analytics dashboard.

Task 14 – Demonstrating End-to-End Big Data Pipeline

Objective: Showcase an integrated big data workflow combining batch processing, real-time streaming, and visualization.

Big Data Analytics – Assignment 1

- Themes:
  - Financial Data Analytics
  - Environment Data Analytics
  - Health Data Analytics

Overview of Assignments

- Three Core Themes:
- 1. Financial Data Analytics (FDA)
- 2. Environment Data Analytics (EDA)
- 3. Health Data Analytics (HDA)
- Objective: Apply Big Data techniques for extraction, processing, and ML.
- Approach: Data Acquisition → Feature Engineering → Modeling → Automation

Theme 1: Financial Data Analytics

- **Scope:** Analyze Securities and Exchange Commission (SEC) filings, financial statements, and market data.
- **Key Tasks:** Corporate Social Responsibility (CSR) analysis, fraud detection, peer benchmarking, sentiment analysis, Mergers and Acquisitions (M&A), and credit rating prediction.
- **Techniques:** Securities and Exchange Commission Electronic Data Gathering, Analysis, and Retrieval (SEC EDGAR), API, NLP on disclosures, statistical ML models.

Theme 2: Environment Data Analytics

- **Scope:** Monitor and analyze environmental data on climate, emissions, and sustainability.
- **Key Tasks:** Carbon footprint tracking, power plant and transport emissions monitoring, deforestation analysis, renewable energy assessment, air quality–health correlation, and environmental policy impact evaluation.
- **Techniques:** Application Programming Interfaces (APIs) such as Carbon Disclosure Project (CDP), Open Air Quality (OpenAQ), National Aeronautics and Space Administration (NASA), and Food and Agriculture Organization (FAO); web scraping; geospatial analysis; and forecasting.

Theme 3: Health Data Analytics

- **Scope:** Use multimodal biomedical, behavioral, and social media data for mental health analysis.
- **Key Tasks:** Electroencephalography (EEG), functional Magnetic Resonance Imaging (fMRI), speech, and wearable data analysis for depression, anxiety, and Post-Traumatic Stress Disorder (PTSD); social media NLP; multimodal fusion; and severity monitoring.
- **Techniques:** Signal processing, NLP, brain connectivity analysis, and multimodal Machine Learning.

Common Workflow Across Themes

Deliverable 1

- Data Acquisition: APIs, public datasets, scraping.
- Preprocessing: Cleaning, normalization, multimodal alignment.
- Feature Engineering: Data and Domain-specific metrics extraction.

Deliverable 2

- Modeling: ML/DL algorithms for prediction, classification, clustering.
- Evaluation: Accuracy, precision, recall, domain KPIs.
- Automation: Pipelines for periodic or real-time updates.

Expected Learning Outcomes

- Mastery in domain-specific big data pipelines.
- Experience with APIs, scraping, and large datasets.
- Application of advanced ML and NLP models.
- Integration of multimodal data sources.
- Development of automated and scalable analytics systems.

Thank you