

Introduction to Big Data Analytics

Introduction to Big Data Analytics

Key concepts, workflow, tools, and real-world applications

1. What Is “Big Data”?

Dimension	Description	Typical Scale
Volume	Massive amount of data generated every second	Terabytes ? Exabytes
Velocity	Speed at which data is created, collected, and processed	Real-time streams, high-frequency logs
Variety	Different formats and sources (structured, semi-structured, unstructured)	Text, images, video, sensor data, logs
Veracity	Trustworthiness & quality of data	Noise, missing values, bias
Value	Potential business or scientific insight extracted	Revenue growth, risk reduction, scientific discovery

The “3Vs” (Volume, Velocity, Variety) are the classic definition; many frameworks add Veracity and Value.

2. Why Analytics on Big Data?

- **Competitive advantage** – faster, data-driven decisions.
 - **Operational efficiency** – predictive maintenance, demand forecasting.
 - **Customer insight** – personalization, churn prediction.
 - **Risk management** – fraud detection, anomaly spotting.
 - **Innovation** – new products/services based on hidden patterns.
-

3. Core Components of a Big Data Analytics Pipeline

1. Data Ingestion

- Batch: Hadoop Distributed File System (HDFS), Amazon S3, Azure Blob.
- Streaming: Apache Kafka, Amazon Kinesis, Azure Event Hubs.

2. Data Storage

Introduction to Big Data Analytics

- Distributed file systems (HDFS, Ceph).
- NoSQL databases (Cassandra, HBase, MongoDB).
- Data warehouses (Snowflake, Redshift, BigQuery).

3. Data Processing

- **Batch processing:** Apache Spark, Hadoop MapReduce, Flink (batch mode).
- **Stream processing:** Spark Structured Streaming, Flink, Apache Storm, Kafka Streams.

4. Data Management & Governance

- Metadata catalogs (AWS Glue, Apache Atlas).
- Data quality tools (Great Expectations, Deequ).
- Security & compliance (encryption, IAM, GDPR/CCPA).

5. Analytics & Machine Learning

- Descriptive: dashboards, SQL queries.
- Diagnostic: correlation, root?cause analysis.
- Predictive: regression, classification, time?series forecasting.
- Prescriptive: optimization, reinforcement learning.

6. Visualization & Reporting

- BI tools: Tableau, Power BI, Looker, Superset.
- Custom notebooks: Jupyter, Zeppelin.

7. Operationalization

- Model serving (MLflow, Sagemaker, KFServing).
- Monitoring & A/B testing.

4. Common Analytical Techniques

Technique	Typical Use Cases	Tools / Libraries
Descriptive statistics	Summarize sales, traffic	Pandas, R, SQL
Clustering (k?means, DBSCAN)	Customer segmentation	Spark MLlib, Scikit?learn
Classification (logistic regression, XGBoost)	Fraud detection, churn prediction	Spark ML, H2O, LightGBM
Regression (linear, ridge, LASSO)	Demand forecasting	Statsmodels, Spark ML
Time?series analysis (ARIMA, Prophet)	Sensor data, stock prices	Facebook Prophet, PyFlux

Introduction to Big Data Analytics

Technique

Typical Use Cases

Tools / Libraries

Anomaly detection

Network intrusion, equipment failure

Isolation Forest, Azure Anomaly Detector

Graph analytics

Social network influence, recommendation

Neo4j, GraphX,