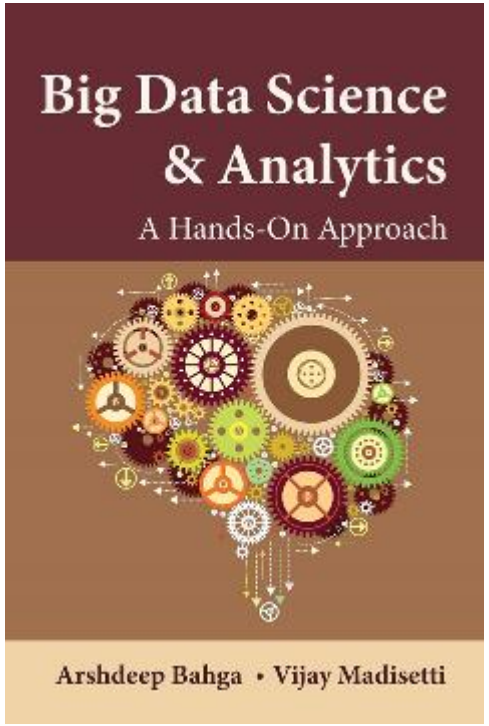
Introduction to Big Data

Prof. Gheith Abandah

Reference

- Chapter 1: **Introduction to Big Data**



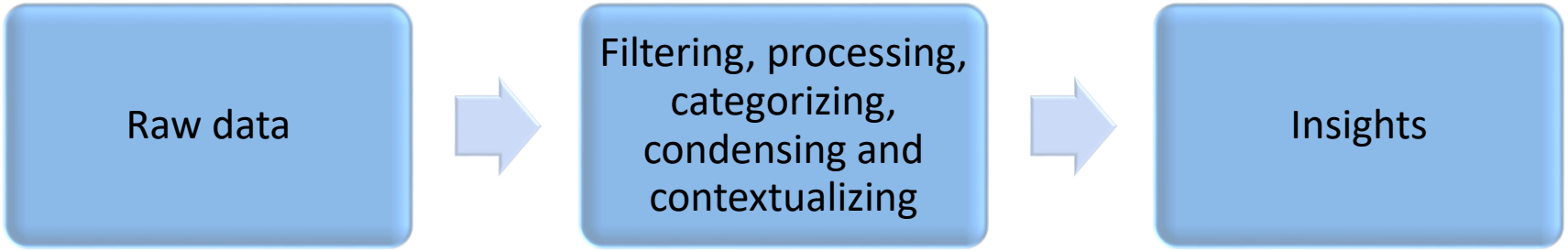
- Arshdeep Bahga and Vijay Madisetti, **Big Data Science and Analytics: A Hands-On Approach**, 2019.
  - Web site: http://www.hands-on-books-series.com/

Outline

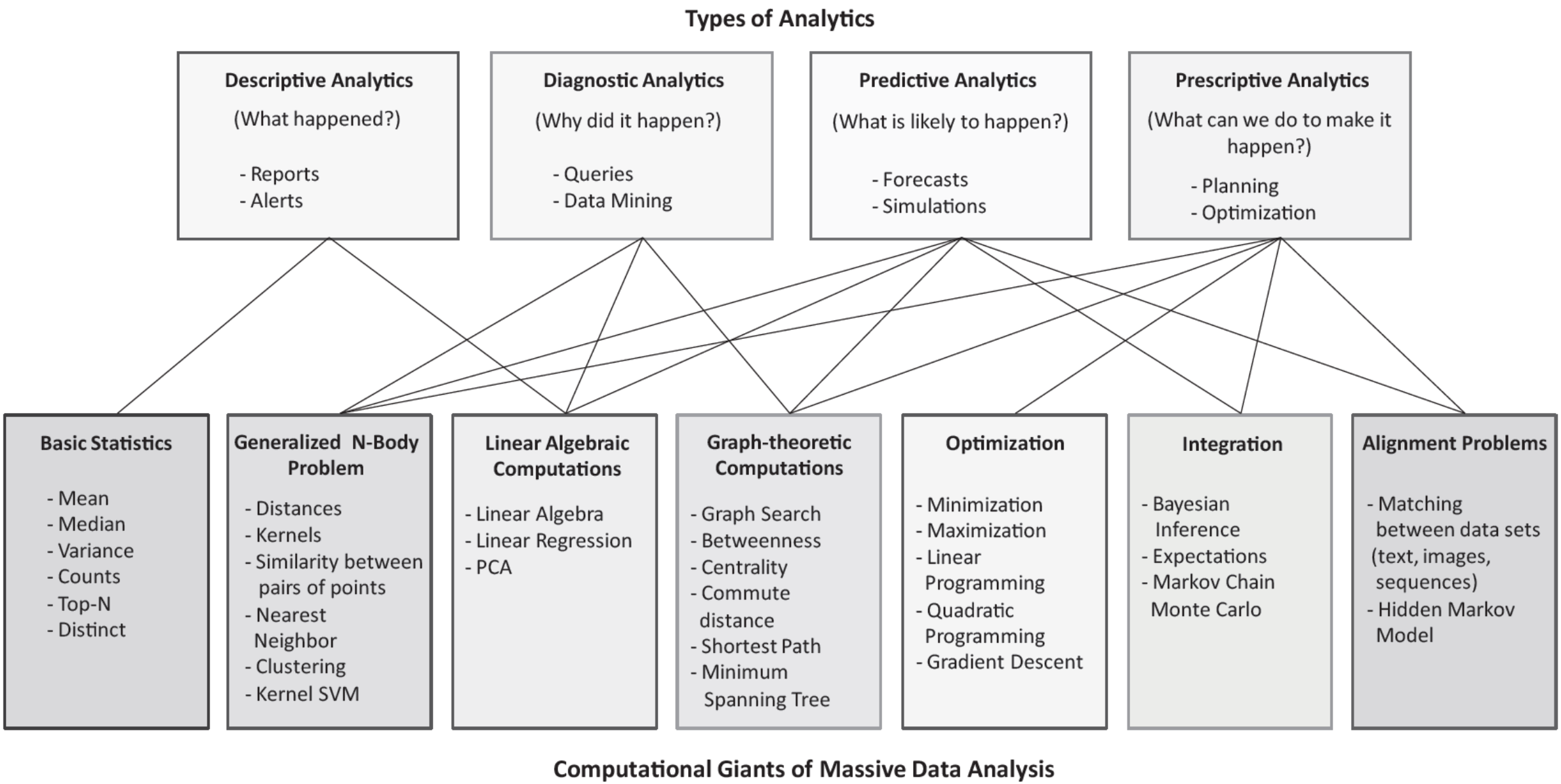
- What is Analytics?
- What is Big Data?
- Characteristics of Big Data
- Domain Specific Examples of Big Data
- Analytics Flow for Big Data
- Big Data Stack
- Mapping Analytics Flow to Big Data Stack
- Case Study: Weather Data Analysis
- Analytics Patterns

What is Analytics?

- Processes, technologies, frameworks and algorithms to extract meaningful **insights from data**.



- **Goals** of the analytics task:
  - To **predict** something
  - To find **patterns** in the data
  - To find **relationships** in the data


Types of Analytics
Descriptive Analytics
(What happened?)
- Reports
- Alerts
Diagnostic Analytics
(Why did it happen?)
- Queries
- Data Mining
Predictive Analytics
(What is likely to happen?)
- Forecasts
- Simulations
Prescriptive Analytics
(What can we do to make it happen?)
- Planning
- Optimization
Basic Statistics
- Mean
- Median
- Variance
- Counts
- Top-N
- Distinct
Generalized N-Body Problem
- Distances
- Kernels
- Similarity between pairs of points
- Nearest Neighbor
- Clustering
- Kernel SVM
Linear Algebraic Computations
- Linear Algebra
- Linear Regression
- PCA
Graph-theoretic Computations
- Graph Search
- Betweenness
- Centrality
- Commute distance
- Shortest Path
- Minimum Spanning Tree
Optimization
- Minimization
- Maximization
- Linear Programming
- Quadratic Programming
- Gradient Descent
Integration
- Bayesian Inference
- Expectations
- Markov Chain Monte Carlo
Alignment Problems
- Matching between data sets (text, images, sequences)
- Hidden Markov Model
Computational Giants of Massive Data Analysis

Outline

- What is Analytics?
- What is Big Data?
- Characteristics of Big Data
- Domain Specific Examples of Big Data
- Analytics Flow for Big Data
- Big Data Stack
- Mapping Analytics Flow to Big Data Stack
- Case Study: Weather Data Analysis
- Analytics Patterns

What is Big Data?

- Big data is defined as collections of datasets whose **volume**, **velocity** or **variety** is **so large** that it is difficult to store, manage, process and analyze the data using traditional databases and data processing tools.
- **Every minutes**:
  - **Facebook** users share nearly **4.16 million** pieces of content
  - **Twitter** users send nearly **300,000** tweets
  - **Instagram** users like nearly **1.73 million** photos
  - **YouTube** users upload **300 hours** of new video content
  - **Apple** users download nearly **51,000** apps
  - **Skype** users make nearly **110,000** new calls
  - **Amazon** receives **4300** new visitors
  - **Uber** passengers take **694** rides
  - **Netflix** subscribers stream nearly **77,000** hours of video

What is Big Data?

- **Big data analytics** deals with **collection**, **storage**, **processing** and **analysis** of this massive-scale data.
- **Specialized tools** and frameworks are required for big data analysis.
- Big data tools and frameworks have **distributed** and **parallel processing architectures** and can leverage the storage and computational resources of a large cluster of machines.
- Big data analytics involves several **steps**:
  - data **cleansing**
  - data **munging** (or wrangling)
  - data **processing** and **visualization**

What is Big Data?

- Some **examples** of big data are listed as follows:
  - **Data** generated by **social networks** including text, images, audio and video data
  - **Click-stream data** generated by web applications such as e-Commerce to analyze user behavior
  - **Machine sensor data** collected from sensors embedded in industrial and energy systems for monitoring their health and detecting failures
  - **Healthcare data** collected in electronic health record (EHR) systems
  - **Logs** generated by web applications
  - **Stock markets data**
  - **Transactional data** generated by banking and financial applications

Outline

- What is Analytics?
- What is Big Data?
- Characteristics of Big Data
- Domain Specific Examples of Big Data
- Analytics Flow for Big Data
- Big Data Stack
- Mapping Analytics Flow to Big Data Stack
- Case Study: Weather Data Analysis
- Analytics Patterns

Characteristics of Big Data

1. **Volume**: Big data is a form of data whose volume is so large that it would not fit on a single machine
2. **Velocity**: Data arrives at very high velocities.
3. **Variety**: Big data comes in different forms such as structured, unstructured or semi-structured, including text data, image, audio, video and sensor data.
4. **Veracity**: Veracity refers to how accurate is the data. To extract value from the data, the data needs to be cleaned to remove noise.
5. **Value**: Refers to the usefulness of data for the intended purpose.

Outline

- What is Analytics?
- What is Big Data?
- Characteristics of Big Data
- Domain Specific Examples of Big Data
- Analytics Flow for Big Data
- Big Data Stack
- Mapping Analytics Flow to Big Data Stack
- Case Study: Weather Data Analysis
- Analytics Patterns

Domain Specific Examples of Big Data

1. **Web**
   - Web analytics
   - Performance monitoring
   - Ad targeting & analytics
   - Content recommendations

2. **Financial**
   - Credit risk modeling
   - Fraud detection

3. **Healthcare**
   - Epidemiological surveillance
   - Patient similarity-based decision intelligence application
   - Adverse drug events prediction
   - Detecting claim anomalies
   - Evidence-based medicine
   - Real-time health monitoring

Domain Specific Examples of Big Data

4. Internet of Things

- Intrusion detection
- Smart parking
- Smart roads
- Structural health monitoring
- Smart irrigation

5. Environment

- Weather monitoring
- Air pollution monitoring
- Noise pollution monitoring
- Forest fire detection
- River floods detection
- Water quality monitoring

Domain Specific Examples of Big Data

6. Logistics & Transportation

- Real-time fleet tracking
- Shipment monitoring
- Remote vehicle diagnostics
- Route generation & scheduling
- Hyper-local delivery
- Cab/taxi aggregators

7. Industry

- Machine diagnosis & prognosis
- Risk analysis of industrial operations
- Production planning and control

Domain Specific Examples of Big Data

8. Retail

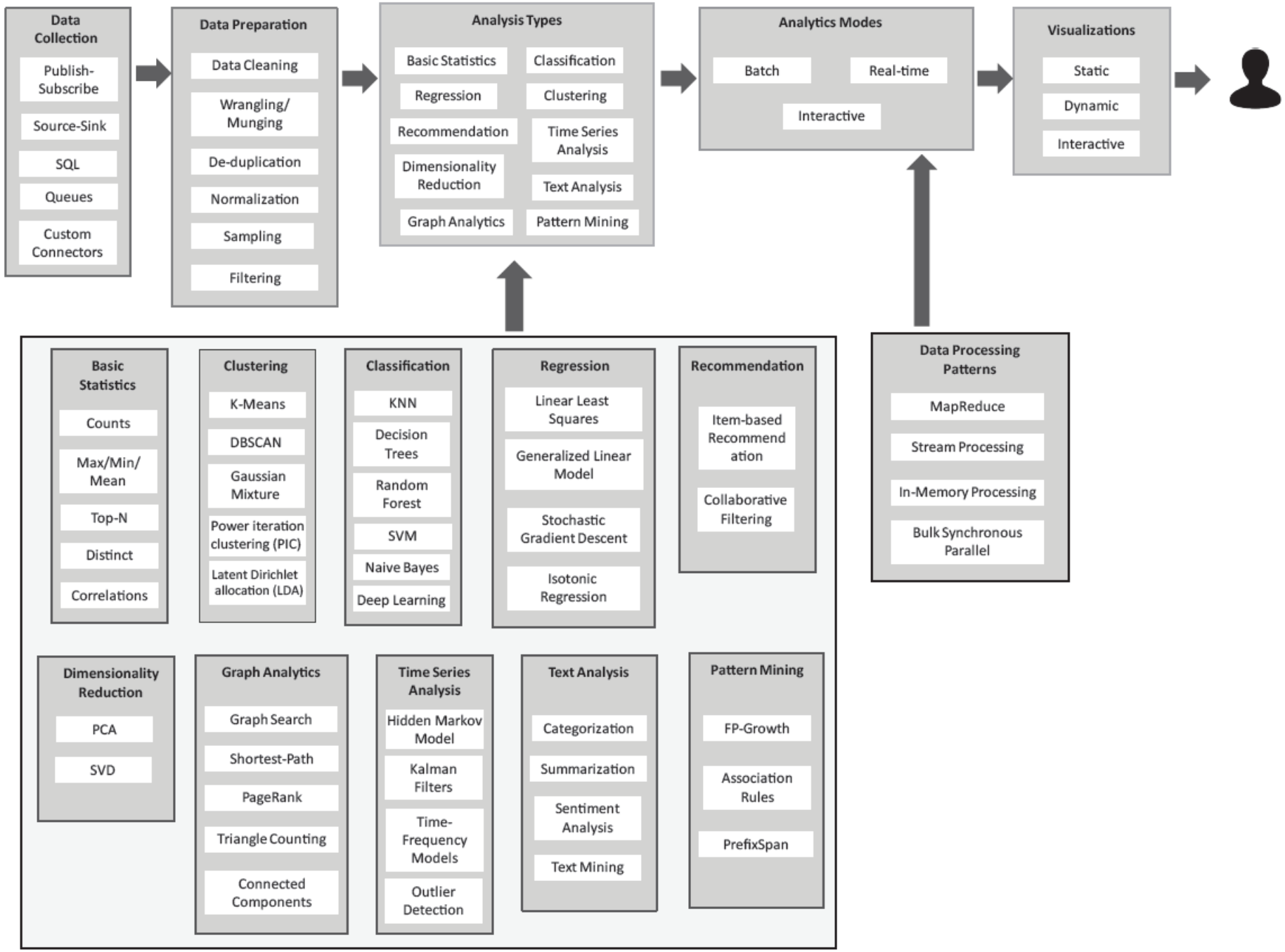
- Inventory management
- Customer recommendations
- Store layout optimization
- Forecasting demand

Outline

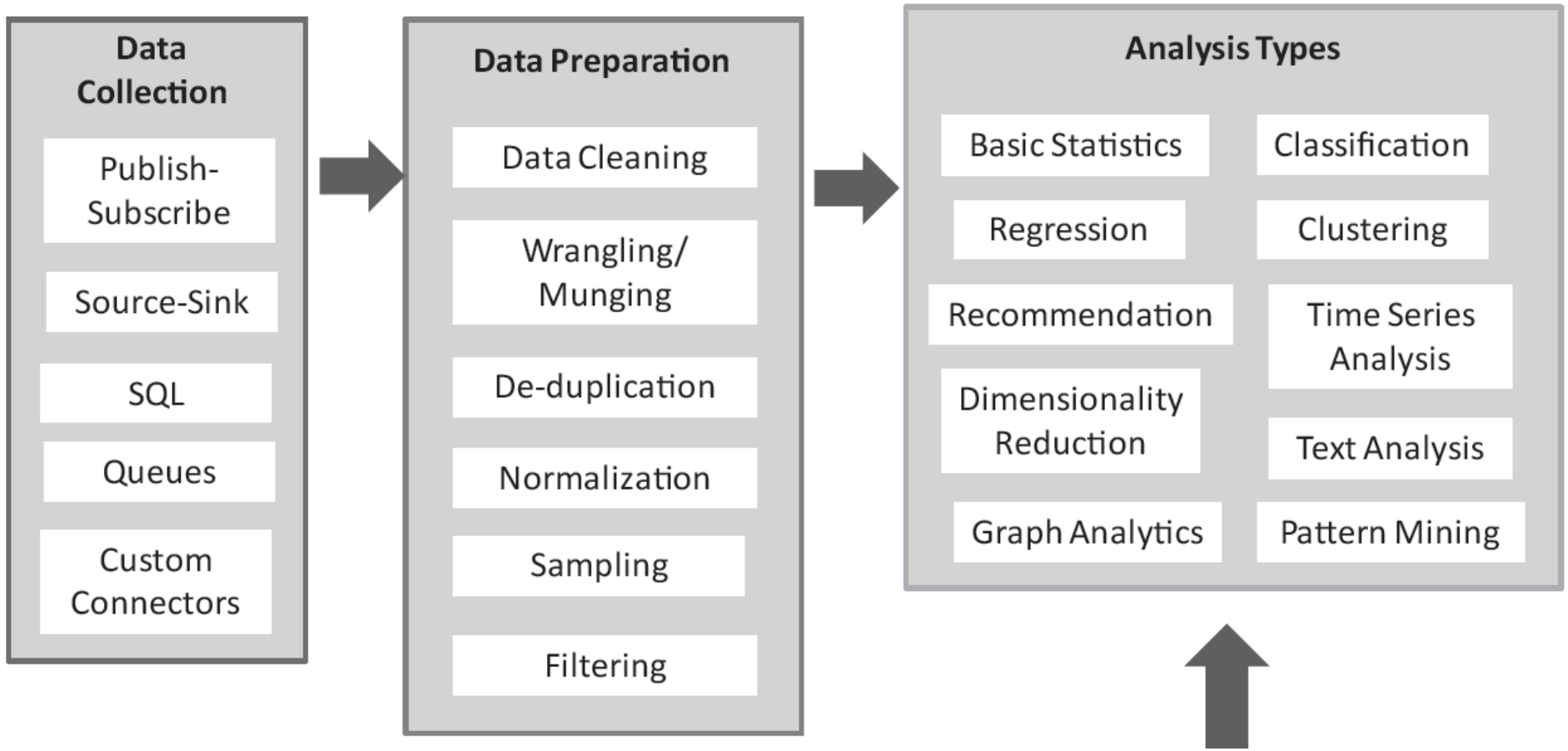
- What is Analytics?
- What is Big Data?
- Characteristics of Big Data
- Domain Specific Examples of Big Data
- Analytics Flow for Big Data
- Big Data Stack
- Mapping Analytics Flow to Big Data Stack
- Case Study: Weather Data Analysis
- Analytics Patterns

Analytics Flow for Big Data

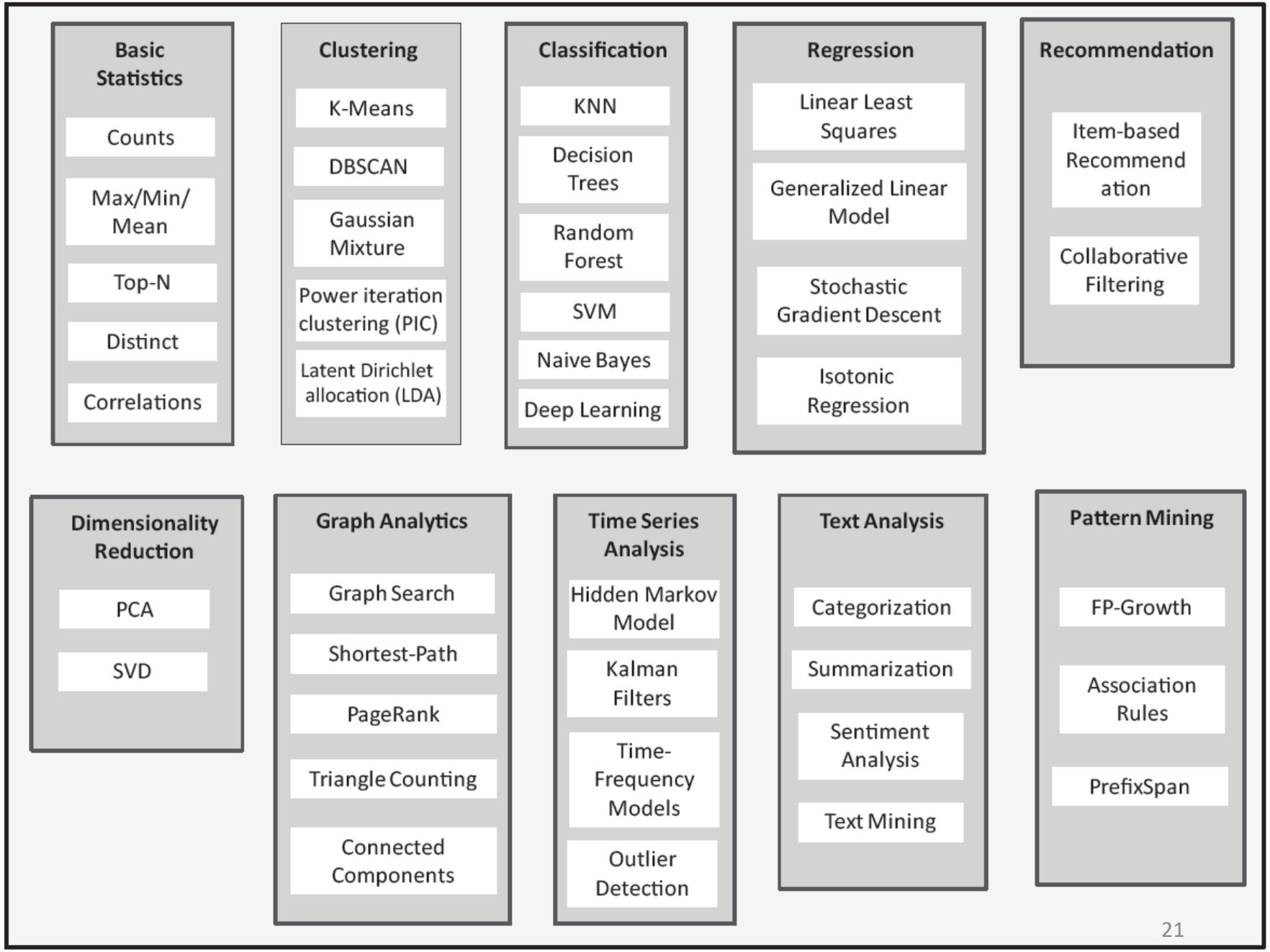
1. **Data Collection**
2. **Data Preparation**
   - Corrupt records, missing values, duplicates, inconsistent abbreviations, inconsistent units, typos and incorrect spellings, incorrect formatting
3. **Analysis**
4. **Visualization**


Data Collection
Publish-Subscribe
Source-Sink
SQL
Queues
Custom Connectors
Data Preparation
Data Cleaning
Wrangling/ Munging
De-duplication
Normalization
Sampling
Filtering
Analysis Types
Basic Statistics
Classification
Regression
Clustering
Recommendation
Time Series Analysis
Dimensionality Reduction
Text Analysis
Graph Analytics
Pattern Mining
Analytics Modes
Batch
Real-time
Interactive
Visualizations
Static
Dynamic
Interactive
Basic Statistics
Counts
Max/Min/ Mean
Top-N
Distinct
Correlations
Clustering
K-Means
DBSCAN
Gaussian Mixture
Power iteration clustering (PIC)
Latent Dirichlet allocation (LDA)
Classification
KNN
Decision Trees
Random Forest
SVM
Naive Bayes
Deep Learning
Regression
Linear Least Squares
Generalized Linear Model
Stochastic Gradient Descent
Isotonic Regression
Recommendation
Item-based Recommend ation
Collaborative Filtering
Data Processing Patterns
MapReduce
Stream Processing
In-Memory Processing
Bulk Synchronous Parallel
Dimensionality Reduction
PCA
SVD
Graph Analytics
Graph Search
Shortest-Path
PageRank
Triangle Counting
Connected Components
Time Series Analysis
Hidden Markov Model
Kalman Filters
Time-Frequency Models
Outlier Detection
Text Analysis
Categorization
Summarization
Sentiment Analysis
Text Mining
Pattern Mining
FP-Growth
Association Rules
PrefixSpan

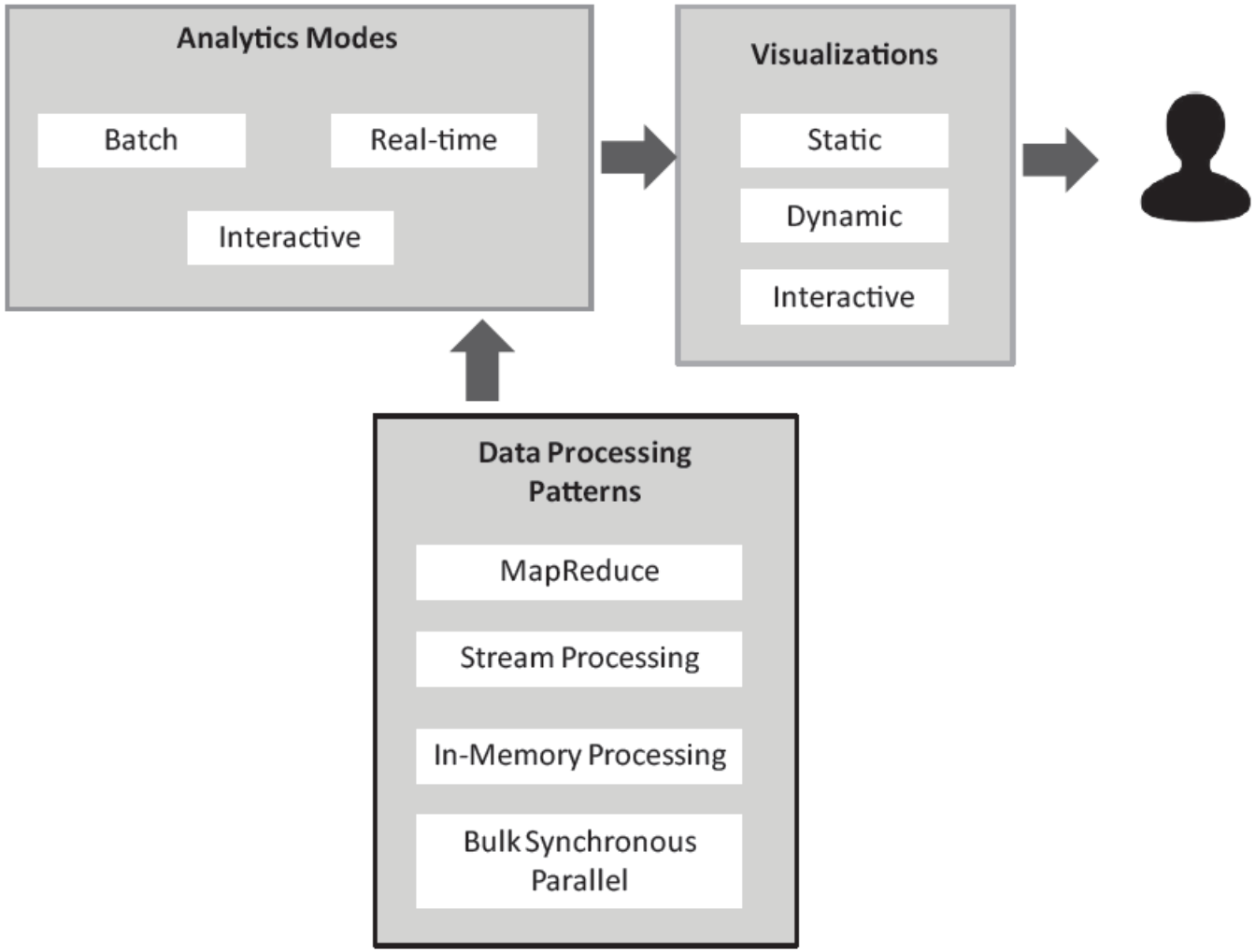
Data Collection, Preparation and Analysis



Analysis Types



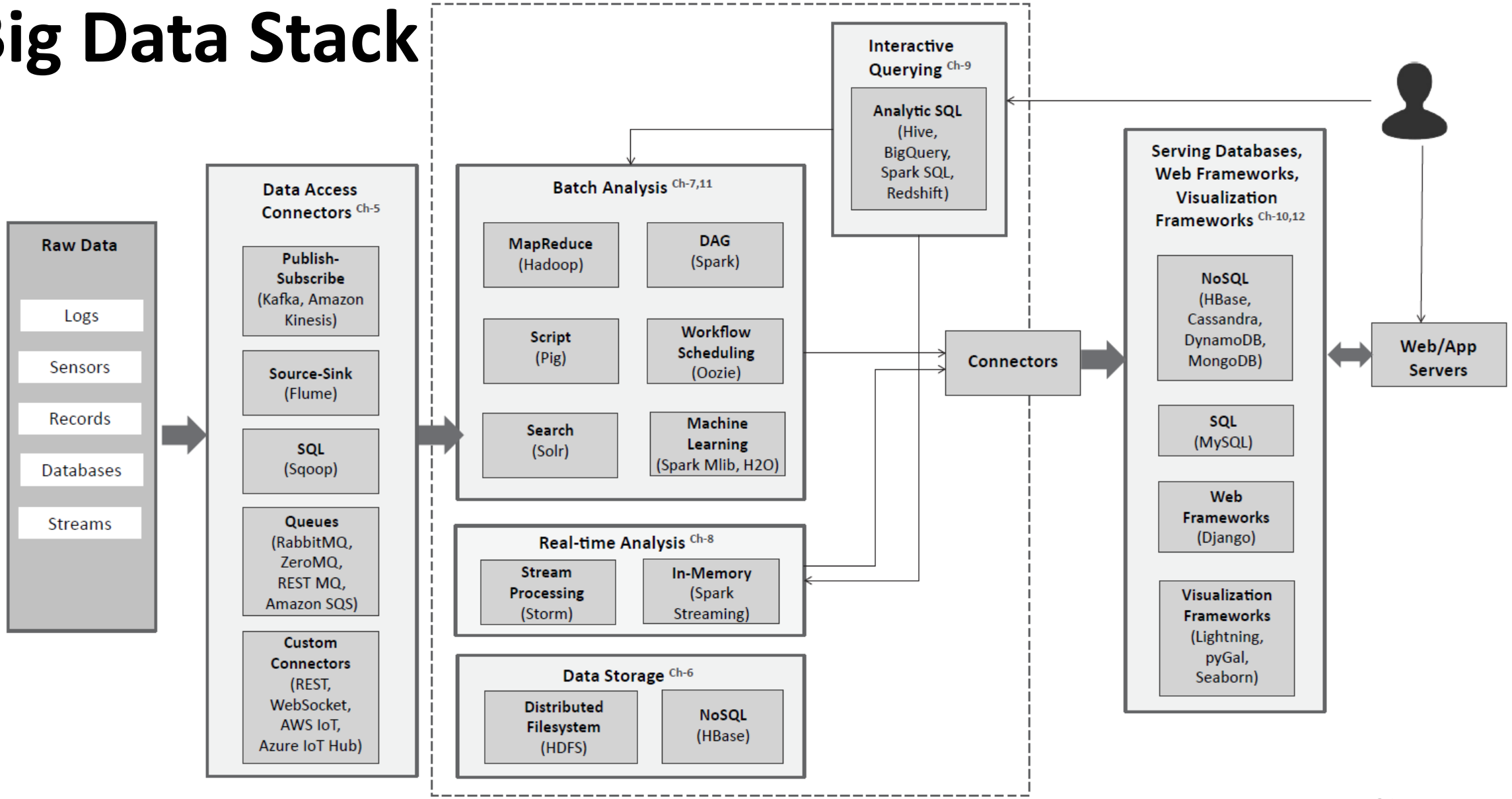
Analytics and Visualization Modes



Outline

- What is Analytics?
- What is Big Data?
- Characteristics of Big Data
- Domain Specific Examples of Big Data
- Analytics Flow for Big Data
- Big Data Stack
- Mapping Analytics Flow to Big Data Stack
- Case Study: Weather Data Analysis
- Analytics Patterns

Big Data Stack



1. Raw Data Sources

1. **Logs** generated by web applications and servers for performance monitoring
2. **Transactional data** generated by applications such as eCommerce, banking and financial
3. **Social media data** generated by social media platforms
4. **Databases**: structured data residing in relational databases
5. **Sensor data** generated by Internet of Things (IoT) systems
6. **Clickstream data** generated by web applications which can be used to analyze browsing patterns of the users
7. **Surveillance data**: Sensor, image and video data generated by surveillance systems
8. **Healthcare data** generated by Electronic Health Record (EHR) and other healthcare applications
9. **Network data** generated by network devices such as routers and firewalls

2. Data Access Connectors

1. **Publish-Subscribe Messaging** is a communication model that involves publishers, brokers and consumers. E.g., Apache Kafka and Amazon Kinesis.
2. **Source-Sink Connectors** allow efficiently collecting, aggregating and moving data from various sources into a centralized data. E.g., Apache Flume.
3. **Database Connectors** can be used for importing data from relational database management systems into big data storage. E.g., Apache Sqoop.

2. Data Access Connectors

4. **Messaging Queues** are useful for push-pull messaging where the producers push data to the queues and the consumers pull the data from the queues. E.g., RabbitMQ, ZeroMQ, RestMQ and Amazon SQS.
5. **Custom Connectors** can be built based on the source of the data and the data collection requirements. Some examples of custom connectors include custom connectors for collecting data from social networks, and connectors for Internet of Things (IoT). E.g., REST, WebSocket, MQTT, and IoT connectors such as AWS IoT and Azure IoT Hub.

3. Data Storage

- The data storage block in the big data stack stores the data collected from the raw data sources using the data access connectors
- Includes
  1. **Distributed file systems**, e.g., Hadoop Distributed File System (HDFS)
  2. **Non-relational (NoSQL) databases**

4. Batch Analytics Frameworks

1. **Hadoop-MapReduce** is a framework for distributed batch processing of big data. Its programming model is used to develop batch analysis jobs which are executed in Hadoop clusters.
2. **Pig** is a high-level data processing language which makes it easy for developers to write data analysis scripts which are translated into MapReduce programs by the Pig compiler.
3. **Oozie** is a workflow scheduler system that allows managing Hadoop jobs. With Oozie, you can create workflows which are a collection of actions (such as MapReduce jobs).

4. Batch Analytics Frameworks

4. **Apache Spark** is an open-source cluster computing framework for data analytics. Spark includes various high-level tools for data analysis such as Spark Streaming for streaming jobs, Spark SQL for analysis of structured data, MLlib, and GraphX for graph processing.
5. **Apache Solr** is a scalable and open-source framework for searching data.
6. **Machine Learning**: Spark MLlib is the Spark's machine learning library which provides implementations of various machine learning algorithms. H2O is an open-source predictive analytics framework which provides implementations of various machine learning algorithms.

5. Real-time Analytics Frameworks

1. **Apache Storm** is a framework for distributed and fault-tolerant real-time computation. Storm can be used for real-time processing of streams of data. Storm can consume data from a variety of sources such as publish-subscribe messaging frameworks (such as Kafka or Kinesis), messaging queues (such as RabbitMQ or ZeroMQ) and other custom connectors.
2. **Spark Streaming** is a component of Spark which allows analysis of streaming data such as sensor data, click stream data, and web server logs. The streaming data is ingested and analyzed in micro-batches. Spark Streaming enables scalable, high throughput and fault-tolerant stream processing.

6. Interactive Querying Systems

1. **Spark SQL** enables interactive querying and is useful for querying structured and semi-structured data using SQL-like queries.
2. **Apache Hive** is a data warehousing framework built on top of Hadoop. It provides an SQL-like query language called Hive Query Language, for querying data residing in HDFS.
3. **Amazon Redshift** is a fast, massive-scale managed data warehouse service. It specializes in handling queries on datasets of sizes up to a petabyte or more parallelizing the SQL queries across all resources in the Redshift cluster.
4. **Google BigQuery** is a service for querying massive datasets. It allows querying datasets using SQL-like queries.

7. Serving Databases, Web & Visualization Frameworks

1. **MySQL** is one of the most widely used Relational Database Management System (RDBMS) and is a good choice to be used as a serving database for data analytics applications where the data is structured.
2. **Amazon DynamoDB** is a fully-managed, scalable, high-performance NoSQL database service. It is an excellent choice for a serving database for data analytics applications as it allows storing and retrieving any amount of data and the ability to scale up or down the provisioned throughput.

7. Serving Databases, Web & Visualization Frameworks

3. **Cassandra** is a scalable, highly available, fault tolerant open-source non-relational database system.
4. **MongoDB** is a document oriented non-relational database system. It is powerful, flexible and highly scalable database designed for web applications and is a good choice for a serving database for data analytics applications.
1. **Django** is an open-source web application framework for developing web applications in Python. It is based on the Model-Template-View architecture and provides a separation of the data model from the business rules and the user interface.

7. Serving Databases, Web & Visualization Frameworks

1. **Lightning** is a framework for creating web-based interactive visualizations.
2. **Pygal** is an easy-to-use Python charting library which supports charts of various types.
3. **Seaborn** is a Python visualization library for plotting attractive statistical plots.

Outline

- What is Analytics?
- What is Big Data?
- Characteristics of Big Data
- Domain Specific Examples of Big Data
- Analytics Flow for Big Data
- Big Data Stack
- Mapping Analytics Flow to Big Data Stack
- Case Study: Weather Data Analysis
- Analytics Patterns

Mapping Analytics Flow to Big Data Stack

- **Data collection tasks**
  - If **high-velocity** data is to be ingested at **real-time**, then a distributed **publish-subscribe** messaging framework such as Apache Kafka or Amazon Kinesis can be used.
  - If the data is to ingested in **bulk** (such as log files), then a **source-sink** such as Apache Flume can be used.
  - For ingesting data from **relational databases**, a framework such as Apache Sqoop can be used.
  - **Custom connectors** can be built based on REST, WebSocket or MQTT.

Mapping Analytics Flow to Big Data Stack

Data Collection

Analysis Type	Framework (Mode)
Publish-Subscribe	Kafka, Kinesis
Source-Sink	Flume
SQL	Sqoop
Queues	SQS, RabbitMQ, ZeroMQ, RESTMQ
Custom Connectors	REST, WebSocket, MQTT

Mapping Analytics Flow to Big Data Stack

Data Preparation

Analysis Type	Framework
Data Cleaning	Open Refine
Data Wrangling	Open Refine DataWrangler
De-Duplication	Open Refine, Pig, Hive, Spark SQL
Normalization Sampling, Filtering	MapReduce, Pig, Hive, Spark SQL

Mapping Analytics Flow to Big Data Stack

Basic Statistics

Analysis Type	Framework (Mode)
Counts, Max, Min, Mean, Top-N, Distinct	Hadoop-MapReduce (Batch), Pig (Batch), Spark (Batch), Spark Streaming (Realtime), Spark SQL (Interactive), Hive (Integrative), Storm (Real-time)
Correlations	Hadoop-MapReduce (Batch), Spark Mlib (Batch)

Mapping Analytics Flow to Big Data Stack

Clustering

Analysis Type	Framework (Mode)
K-Means	Hadoop-MapReduce (Batch), Spark Mlib (Batch & Real-time) H2O (Batch)
DBSCAN	Spark (Batch)
Gaussian Mixture	Spark Mlib (Batch)
PIC	Spark Mlib (Batch)
LDA	Spark Mlib (Batch)

Mapping Analytics Flow to Big Data Stack

Classification

Analysis Type	Framework (Mode)
KNN	Spark Mlib (Batch , Realtime)
Decision Trees	Spark Mlib (Batch, Realtime)
Random Forest	Spark Mlib (Batch , Realtime), H2O (Batch)
SVM	Spark Mlib (Batch , Realtime)
Naïve Bayes	Spark Mlib (Batch, Realtime), H2O (Batch)
Deep Learning	H2O (Batch)

Mapping Analytics Flow to Big Data Stack

Regression

Analysis Type	Framework (Mode)
Linear Least Squares	Spark Mlib (Batch, Realtime)
Generalized Linear Model	H2O (Batch)
Stochastic Gradient Descent	Spark Mlib (Batch, Realtime)
Isotonic Regression	Spark Mlib (Batch, Realtime)

Mapping Analytics Flow to Big Data Stack

Graph Analytics

Analysis Type	Framework (Mode)
Graph Search	Spark GraphX (Batch)
Shortest-Path	Spark GraphX (Batch)
PageRank	Spark GraphX (Batch)
Triangle Counting	Spark GraphX (Batch)
Connected Components	Spark GraphX (Batch)

Time Series Analysis

Analysis Type	Framework (Mode)
Kalman Filter	Spark (Realtime)
Time Frequency Models	Spark (Realtime)
Outlier Detection	Storm (Realtime), Spark (Batch, Realtime)

Dimensionality Reduction

Analysis Type	Framework (Mode)
SVD	Spark Mlib (Batch)
PCA	Spark Mlib (Batch), H2O (Batch)

Recommendation

Analysis Type	Framework (Mode)
Item-bases Recommendation	Spark Mlib (Batch)
Collaborative Filtering	Spark Mlib (Batch)

Text Analysis

Analysis Type	Framework (Mode)
Categorization	Hadoop-MapReduce (Batch), Storm (Realtime), Spark (Batch, Realtime)
Summarization	Spark (Batch)
Sentiment Analysis	Storm (Realtime), Spark (Batch, Realtime)
Text Mining	Storm (Realtime), Spark (Batch, Realtime)

Pattern Mining

Analysis Type	Framework (Mode)
FP-Growth	Spark Mlib (Batch)
Association Rules	Spark Mlib (Batch)
PrefixSpan	Spark Mlib (Batch)

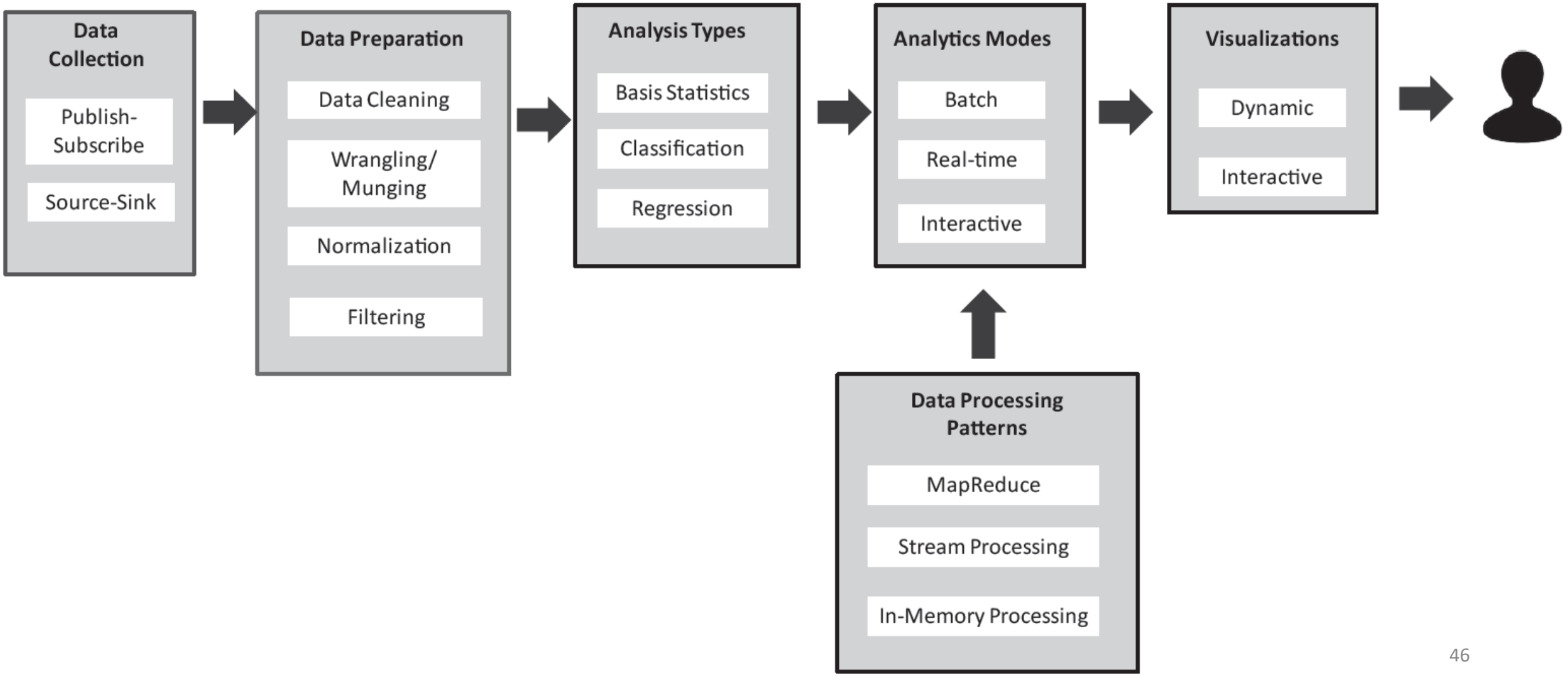
Visualization

Analysis Type	Framework (Mode)
Web Frameworks	Django, Flask
SQL Databases	MySQL
NoSQL Databases	Hbase, DynamoDB, Cassandra, MongoDB
Visualization Frameworks	Lightning, pyGal, Seaborn

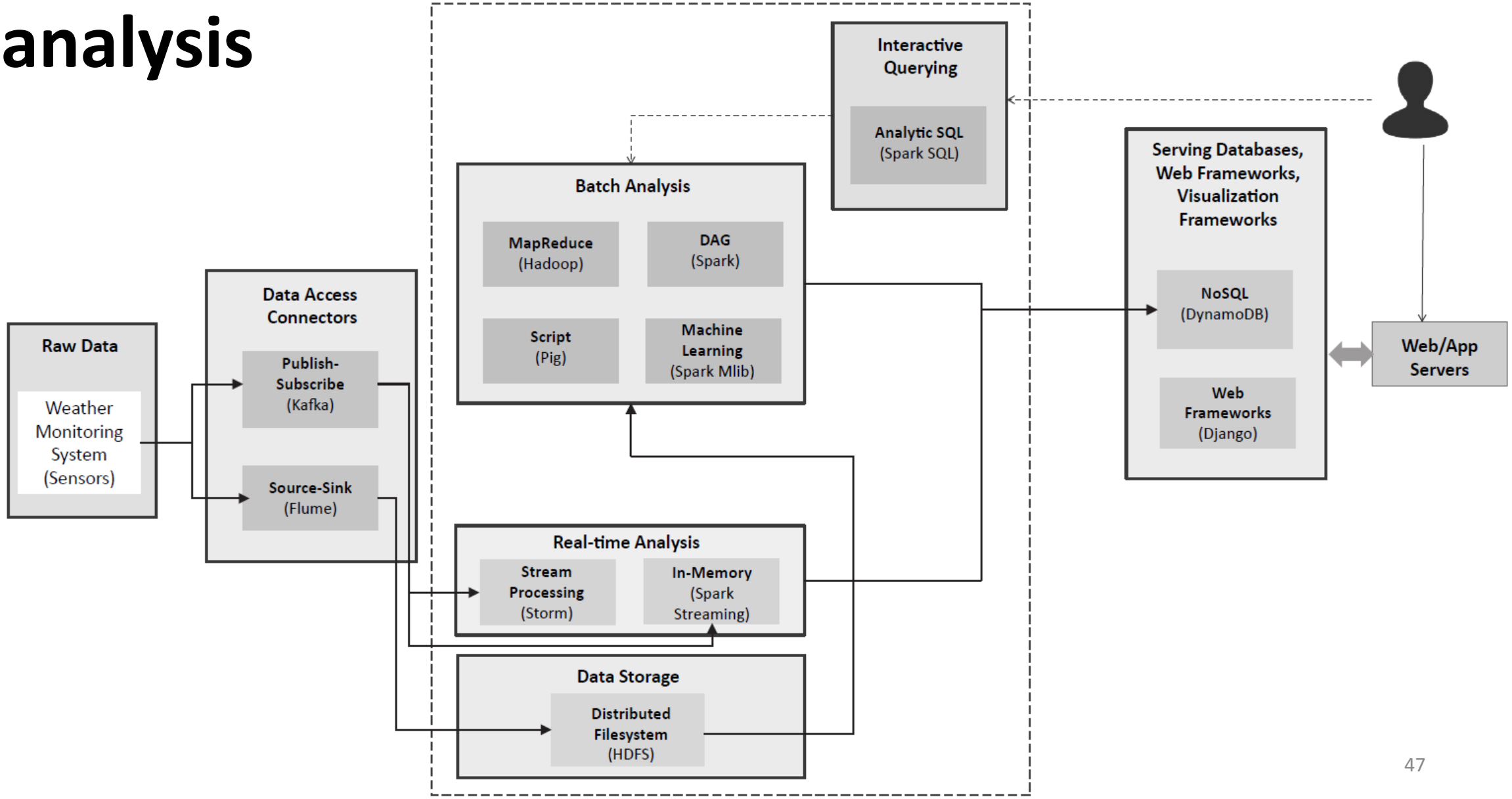
Outline

- What is Analytics?
- What is Big Data?
- Characteristics of Big Data
- Domain Specific Examples of Big Data
- Analytics Flow for Big Data
- Big Data Stack
- Mapping Analytics Flow to Big Data Stack
- Case Study: Weather Data Analysis
- Analytics Patterns

Analytics flow for weather data analysis application



Using Big Data stack for weather data analysis



Outline

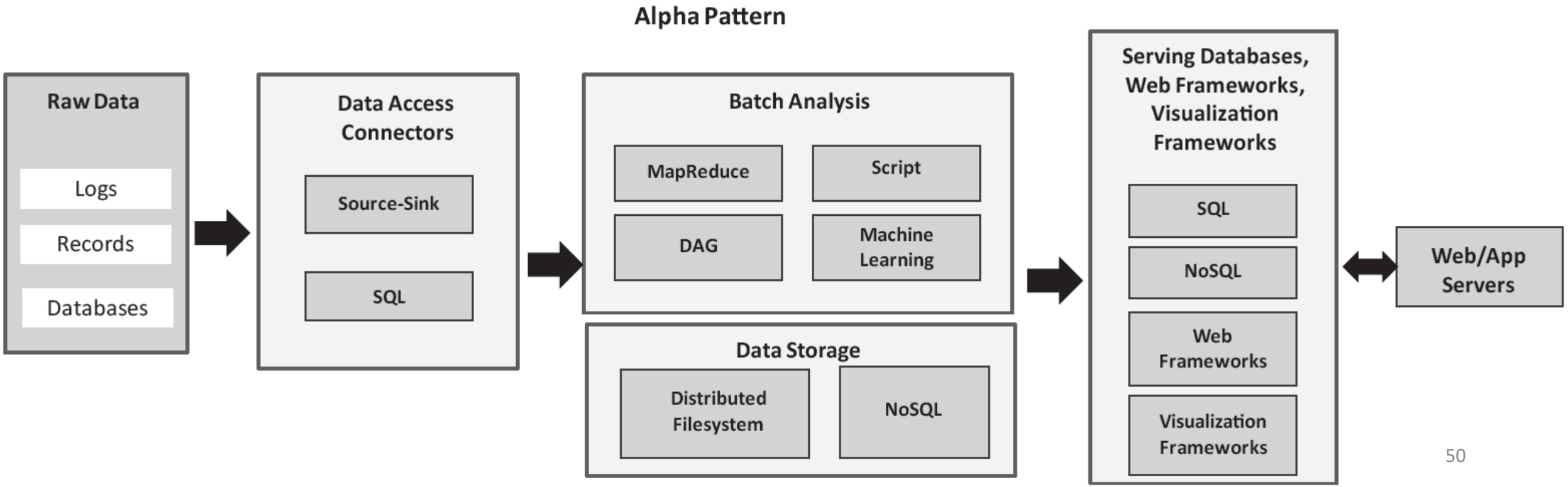
- What is Analytics?
- What is Big Data?
- Characteristics of Big Data
- Domain Specific Examples of Big Data
- Analytics Flow for Big Data
- Big Data Stack
- Mapping Analytics Flow to Big Data Stack
- Case Study: Weather Data Analysis
- Analytics Patterns

Analytics Patterns

- **Four patterns** that are tools and frameworks for:
  - **collecting** and **ingesting** data from various sources into the big data analytics infrastructure, distributed filesystems and non-relational (NoSQL) databases
  - **data storage**
  - **processing** frameworks for batch and real-time analytics
  - **interactive querying**
  - **serving** databases, and web and **visualization**

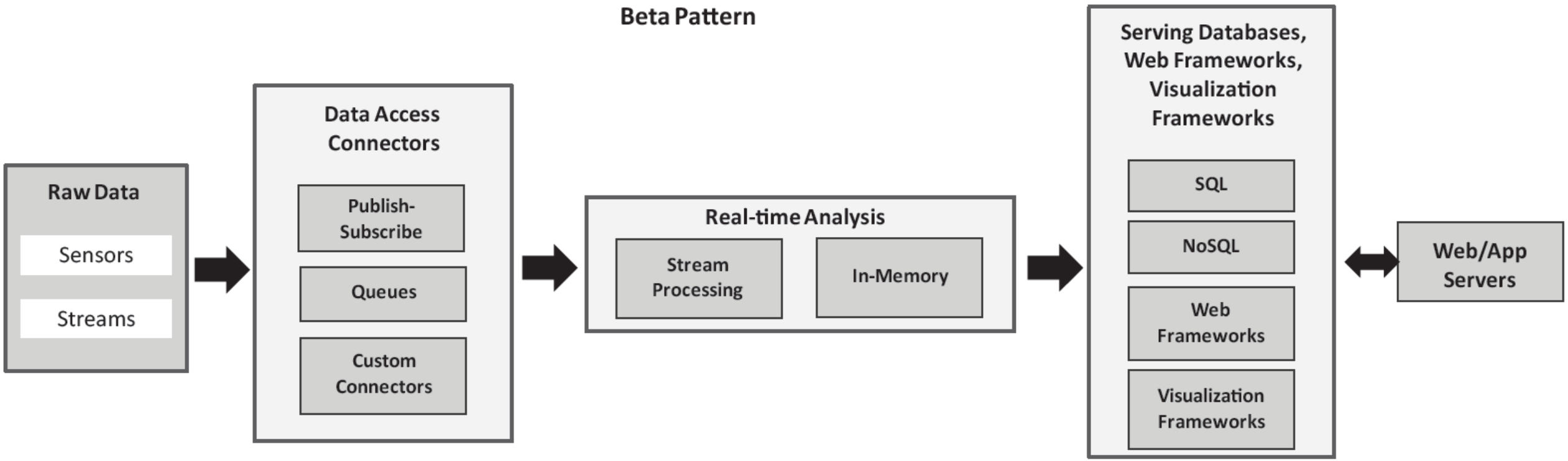
Alpha Pattern

- For ingesting **large volumes** of data into a distributed filesystem
  - Web analytics, weather monitoring, epidemiological surveillance, and machine diagnosis



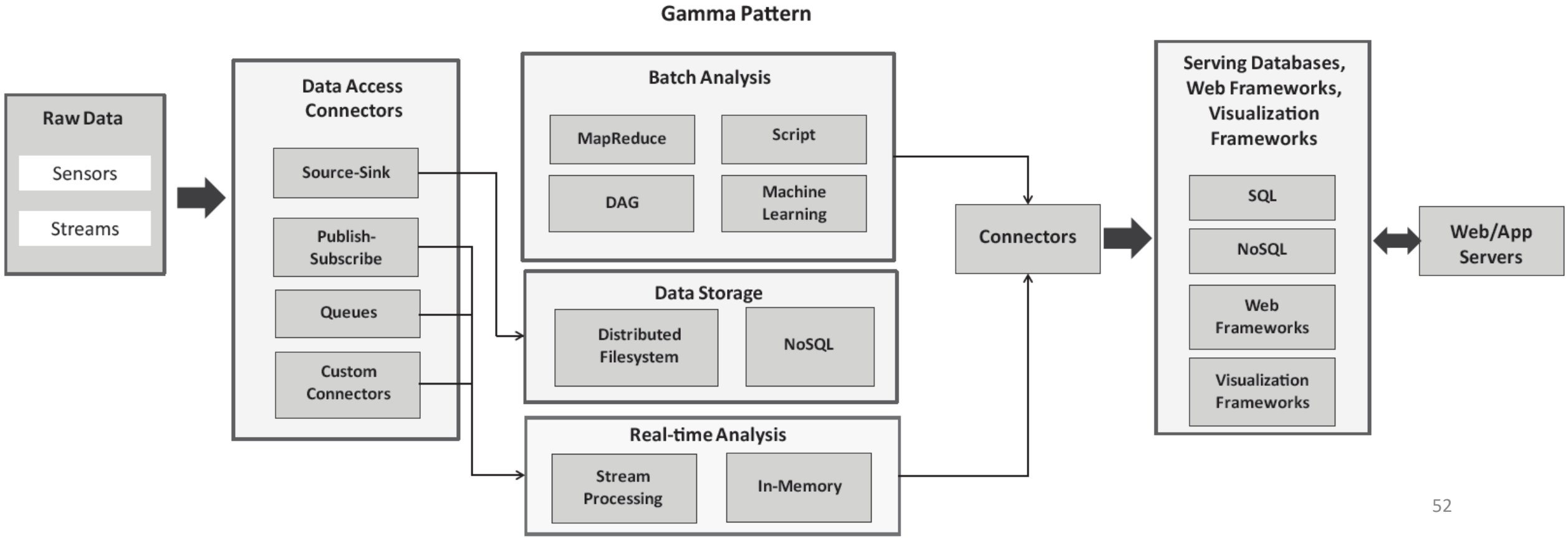
Beta Pattern

- For ingesting **streaming data**
  - Internet of Things applications and real-time monitoring applications



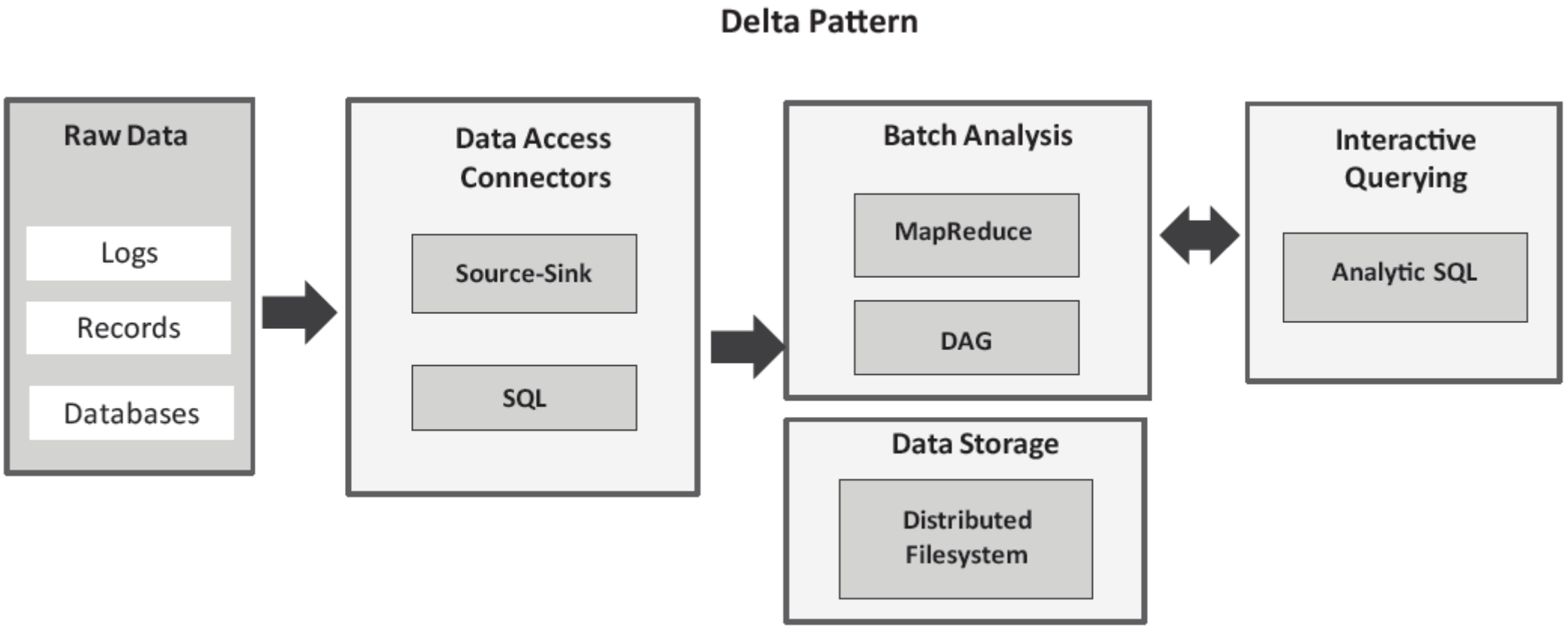
Gamma Pattern

- Combines **batch** and **real-time** analysis patterns
  - large number of IoT devices

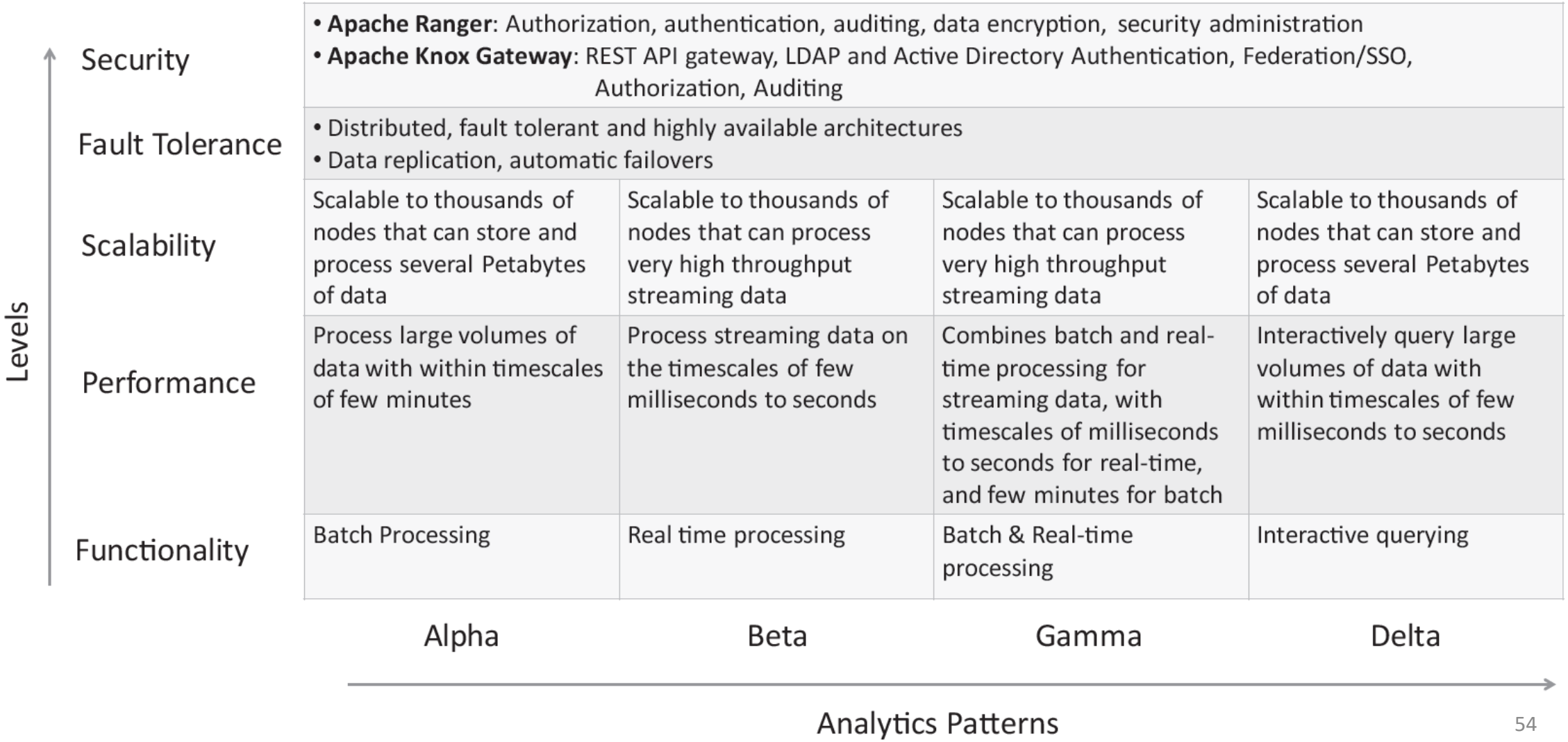


Delta Pattern

- For **interactive querying**
  - web analytics, advertisement targeting, inventory management, production planning and control, and various types of enterprise applications



Complexity levels for analytics patterns



Mapping of blocks in analytics patterns to AWS and Azure services

Data Access Connector	AWS Service	Azure Service
Publish-Subscribe	AWS Kinesis	Azure Event Hubs
Source-Sink	Flume on EMR	Flume on HDInsight
SQL	Sqoop on EMR	Sqoop on HDInsight
Queues	AWS SQS	Azure Queue Service
Custom Connectors	AWS IoT	Azure IoT Hub

Data Storage	AWS Service	Azure Service
Distributed Filesystem	HDFS on AWS EMR	HDFS on Azure HDInsight
NoSQL	AWS DynamoDB	Azure DocumentDB

Batch Analysis	AWS Service	Azure Service
MapReduce	Hadoop on AWS EMR	Hadoop on Azure HDInsight
Script	Pig on AWS EMR	Pig on Azure HDInsight
DAG	Spark on AWS EMR	Spark on Azure HDInsight
Machine Learning	Spark MLlib on AWS EMR, AWS Machine Learning	Spark MLlib on Azure HDInsight, Azure Machine Learning

Real-time Analysis	AWS Service	Azure Service
Stream Processing	Storm cluster on AWS EC2	Storm on Azure HDInsight, Azure Stream Analytics
In-Memory Processing	Spark on AWS EMR	Spark on Azure HDInsight

Serving Databases, Web & Visualization Frameworks	AWS Service	Azure Service
SQL	AWS RDS	Azure SQL DB
NoSQL	AWS DynamoDB	Azure DocumentDB
Web Framework	Django on AWS EC2 instance	Django on Azure Virtual Machines instance
Visualization Framework	Lightning, Pygal, Seaborn, on AWS Virtual Machines instance	Lightning, Pygal, Seaborn, on Azure Virtual Machines instance, Azure Power BI

Summary

- What is Analytics?
- What is Big Data?
- Characteristics of Big Data
- Domain Specific Examples of Big Data
- Analytics Flow for Big Data
- Big Data Stack
- Mapping Analytics Flow to Big Data Stack
- Case Study: Weather Data Analysis
- Analytics Patterns