Data Curation - Tutorial

Never Stand Still

Computer Science and Engineering

Alireza Tabebordbar

Comp 9313- Big Data Management

Data Curation Tutorial

Running Example

In this tutorial:
We extract Tweets from Twitter fire hose, then we perform pre-processing on the extracted Tweets. We also, index the Tweets for fast search and retrieval. Finally, we extract a set of named entities from Tweets to add value to the extracted information.

Goal of the Tutorial:

- We demonstrate how to create a pipeline for curating data.
- We explain how a big data streaming technology such as Apache Kafka can be coupled with a data curation pipeline.
- We explain how data can be transformed for adding values and extracting insight.

Big Data Curation

The data curation tasks we demonstrate in this example:

- **Identifying** relevant data sources
- **Ingesting** data and knowledge
- **Cleaning**
- **Integration**
- **Transformation** (Normalization and aggregation)
- **Adding Value** (Preparing Raw Data for Analytics):
  - **Extraction**
  - **Enrichment**
  - **Linking**
  - **Summarization**

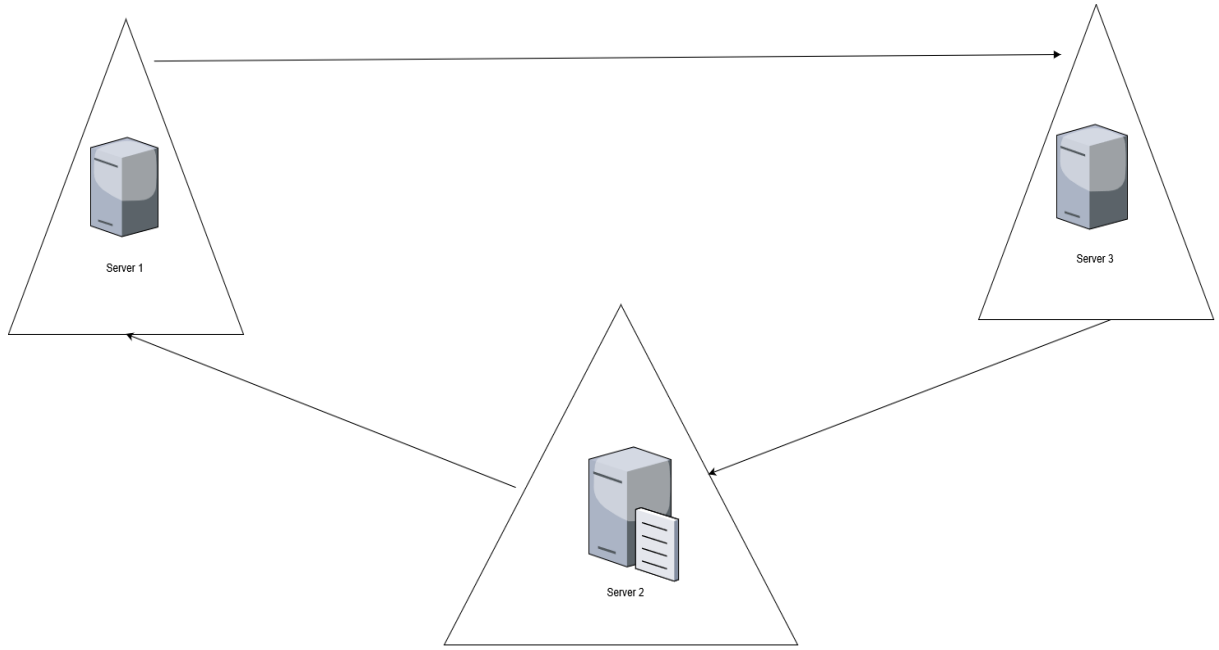
Data Ingestion

Introduction to data streaming and Apache Kafka

Apache Kafka is an open-source stream-processing software platform developed by LinkedIn and donated to the Apache Software Foundation, written in Scala and Java. The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds.

Apache Kafka Characteristics:

- Kafka provides a messaging system that replicates stream of records across a cluster of servers.
- The Kafka cluster stores streams of *records* in categories called *topics*.
- Each record consists of a key, a value, and a timestamp.

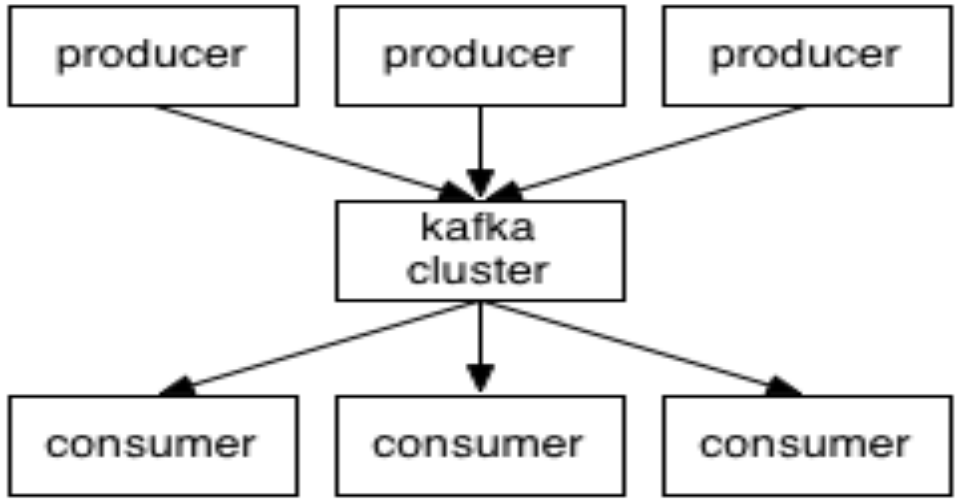


Data Ingestion

Introduction to data streaming and Apache Kafka

Kafka Producer and Consumer:

The two main components of Kafka are **Producer** and **Consumer** APIs:

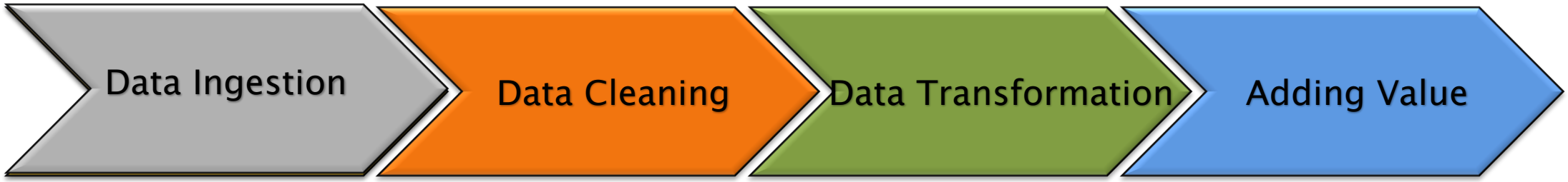


Producer API:

- Allows an application to publish a stream of records to one or more Kafka topics.

Consumer API:

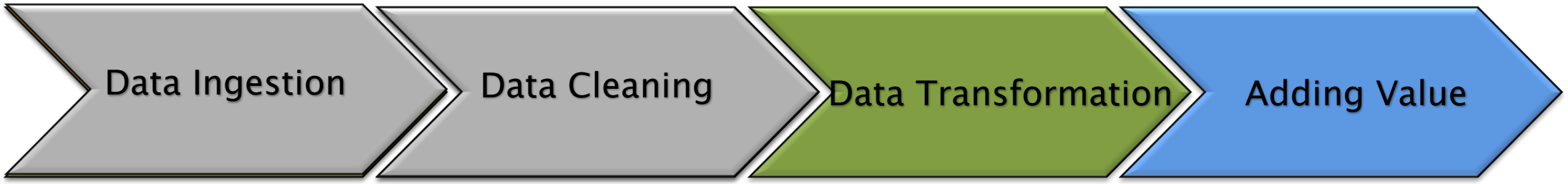
- Allows an application to subscribe to one or more topics and process the stream of records produced to them.



Data Ingestion

Cleaning

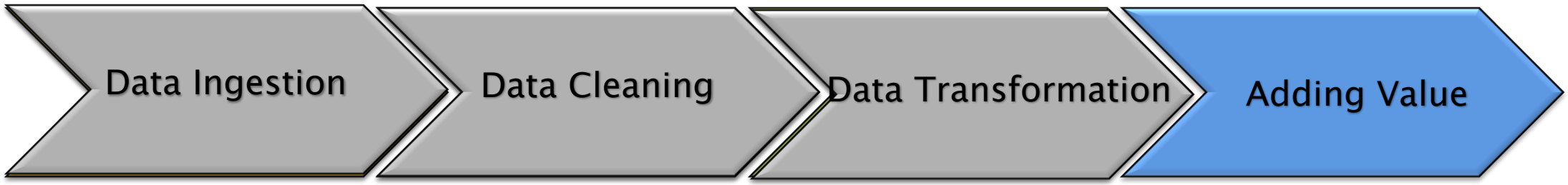
- **In the Cleaning Process, we perform some pre-processing tasks including:**
  - We remove URLs and IMOJI from Tweets
  - We lowercase the Tweets tokens
  - We keep proper names uppercase



Data Transformation

Indexing Data

- **In this step, we create we store the Tweets as a set of Index. Indexing is made up of two parts:**
  - Creation of index
  - Search
- **Indexing allows to create to retrieve and search data much faster compared to conventional retrieval approaches.**



Adding Value

Extracting Named Entities

- **We extract a set of named entities from Tweets text and prints their frequencies.**
- **Named Entities are real-world objects, such as**
  - Persons
  - Locations
  - Organizations
  - Products
  - etc.

that can be denoted with a proper name. It can be abstract or have a physical existence. Examples of named entities include *Barack Obama*, *New York City*, *Volkswagen Golf*, or anything else that can be named.