# Assignment 2: Design and Implementation of Data Architecture and Data Processing Pipelines

| | |
|---|---|
| Deliverables | A report (less than 7 pages) - upload to Canvas.<br>A working application (demo).<br>Source code: a link to a GitHub project should be provided.<br><br>If you are using a private GitHub repository, add me (IndikaKuma) to it. |
| Submission Deadline | 3rd December 23:59 |
| Demonstration | No live demos.<br>The students need to upload a recoded video. You can update the original submission.<br><br>Deadline 5th December 23:59 |
| Late Submission Penalty | 10% reduction (0 < delay < 7 days)<br>50% reduction (delay > 7 days) |
| Grade Percentage | 20% of the course grade (i.e., 20 points from 100 points). |

## Goals

- Design and implement data architectures and data processing pipelines using Apache Spark and GCP services.

## Skills Required

- Ability to create, configure, and use a data architecture in the GCP
- Ability to create, deploy, and execute batch and stream data processing jobs with Apache Spark

## Assignment Description

### Definition of a Data Pipeline

We use the definition from https://martinfowler.com/articles/cd4ml.html#DataPipelines

*"Pipeline is an overloaded term, especially in ML applications. We want to define a "data pipeline" as the process that takes input data through a series of transformation stages, producing data as output. Both the input and output data can be fetched and stored in different locations, such as a database, a stream, a file, etc. The transformation stages are usually defined in code, although some ETL tools allow you to represent them in a graphical form. They can be executed either as a batch job, or as a long-running streaming application."*

Spark program/job defines a data pipeline programmatically (code).

## Definition of a Data Infrastructure

We use the definition from the following article for a data infrastructure. A data infrastructure can execute/host multiple data pipelines.

https://a16z.com/emerging-architectures-for-modern-data-infrastructure/

## Assignment Constraints

- The data architecture (*of a data infrastructure*) must be based on recognized data architectures such as Microsoft Big Data Architecture (one used in the labs), Lambda, Kappa, and Data Warehouse.
- The data architecture must be implemented with the most appropriate tools, such as Spark, BigQuery, Google Cloud Storage, and Kafka.
- **Two separate** data pipelines should be implemented. The pipelines can be batch processing **and/or** data stream processing pipelines. Each pipeline will be a Spark program/job.
  - 2 batch or
  - 2 stream or
  - 1 batch and 1 stream
- The students **cannot** use the datasets used in the labs.
- The students need to develop their own Spark programs. The students **cannot use** the complete Spark programs from the articles or GitHub as-is. For example, the students cannot simply copy and use a Spark program implemented for a specific use case.
- The data pipelines should implement a sufficient number of data processing steps (see the examples given later for the level of complexity expected). Machine learning is **not** the focus of this assignment. While the students can use machine learning in their pipelines, simply training models and making predictions is insufficient. For example, there should be reasonably complex data preprocessing where data pipelines are used. The students are encouraged to focus on batch and stream data processing use cases.
  - An assignment report from the previous batch can be found in Canvas (*Modules >> Assignments*). You can check it to get an idea about the assignment.

## Goals of the Data Pipelines

A data pipeline is designed to support some end-user goals. Each data pipeline should support at least one business-related or end-user goal. The following are three goal examples (**hypothetical**).

**Goal Example 1:** Understand the customer satisfaction trends (for Dutch railway) based on tweets

**Goal Example 2:** *Clean* and *integrate* tweets and stock data, and then *extract* the features that can be used to build a model to understand the correlation between the stock market movements of a company and sentiments in tweets.

**Goal Example 3**: Analyze and produce the statistics (e.g., average car model prices, filtered car model listings, state-level average prices, and manufacturer listings) for the current used cars market in the United States.

Goal Example 3 is from one of the student projects from last year's course. The project used the dataset from https://www.kaggle.com/austinreese/craigslist-carstrucks-data

**Note**: Two data pipelines in combination may implement a single goal.

## Number of Data Datasets

Students can use any number of datasets, including just a single dataset. The students can use any publicly available datasets.

Students can also generate and use synthetic datasets. A good tool for data generation is https://www.mockaroo.com/ or **LLMs**

# Examples of Data Architecture Implementations

Figure 1 shows the data architecture we used in the labs.

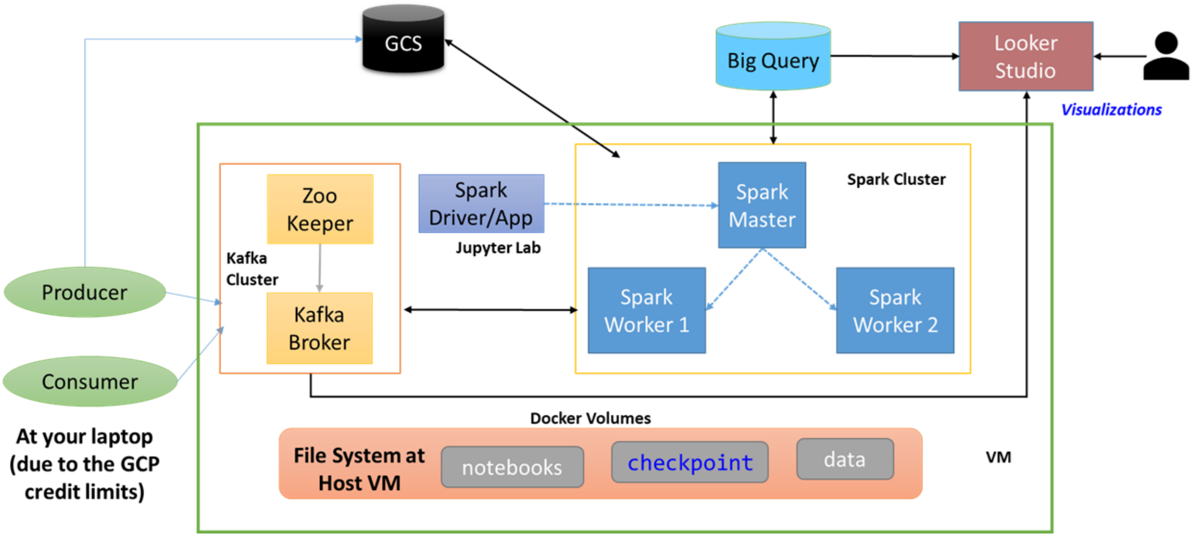

**Figure 1. A Data Architecture Implementation based on Microsoft Big Data Architecture**

Figure 2 shows an instantiation of the Lambda architecture style for an IoT data processing use case. The serving layer can be replaced with BigQuery, and as an alternative to HDFS, the GCS buckets, or local file systems (in a VM) can be used.

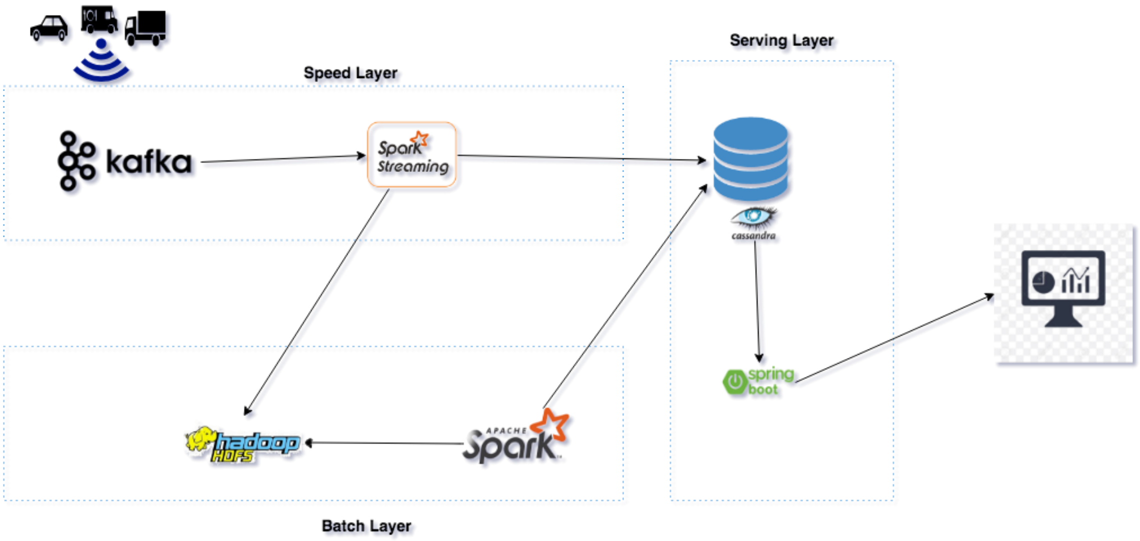

**Figure 2. Lambda Architecture Used by [https://github.com/apssouza22/lambda-arch](https://github.com/apssouza22/lambda-arch)**

# Examples of Data Pipelines

See *Assignment 2 examples from the last batch on* Canvas. Some students used several data pipelines. However, **two** reasonably complex pipelines are good enough for the assignment. A few more use cases were discussed in Lab 6 (live) – see *Lab 6-DE 2025 –Implementing Big Data Architectures with Containers – Live.pptx*

# Report Guidelines

| | |
|---|---|
| Page Limit | Less than 7 pages (excluding front page, references, and appendix) |
| Report Content | 1. Overview of the Data Pipelines<br>What do the data pipelines do? Describe their goals/requirements.<br>2. Design and Implementation of Data Architecture<br>Describe data architecture and its implementation. Use diagrams as necessary.<br>3. Design and Implementation of Data Pipelines<br>Describe data processing pipelines. As necessary, use diagrams, e.g., flow charts, to show the logical design of the data processing pipelines (i.e., processing steps and their flow).<br>4. Reflection on Design and Implementation of Data Architecture and Data Pipelines |

| | |
|---|---|
| | What are the possible alternative designs for your data architecture and data pipelines? First, identify at least one alternative design. Then, compare it (them) with the designs you have implemented in terms of potential strengths and weaknesses.<br>Reflect on the process of implementing your data architecture and pipelines. What were the difficulties you faced? How did you overcome them?<br>5. Individual Contributions of Students<br>Briefly describe the individual contributions of each student in the group.<br>**Technology statement student(s)**<br>For the assignment, you must add the technology statement, using the text below. Replace the text in capital letters with the requested information.<br>During the preparation of this work, I/We used **[NAME TOOL / SERVICE / VERSION OF AI TOOL]** in order to **[REASON].** The following parts of the assignment were affected/generated by AI tool usage: **[INTRODUCTION / METHODS / xxx, DISCUSSION]**. After using this tool/service, **[NAME STUDENT(S)]** evaluated the validity of the tool's outputs, including the sources that generative AI tools have used, and edited the content as needed. As a consequence**, [NAME STUDENT (S)]** take(s) full responsibility for the content of their work. |

# Marking Scheme

The grading considers implementation, demonstration, and report.

| | |
|---|---|
| Data Architecture (design, implementation, and demonstration) | 15% |
| Two Data Processing Pipelines (design, implementation, and demonstration) | 30*2 = 60% |
| Report | 25% |
| Total | 100% |

## AI Policy for Data Engineering Course

During the exam(s) itself, the use of AI is not permitted (**AI - Index - Level 1**).

This assignment (**AI - Index - Level 4**) allows you to use tools such as ChatGPT to complete certain elements of the tasks (e.g., literature search, data analyses, audio transcription). AI serves as a support tool (i.e., like your assistant), but you are WWresponsible for critically evaluating its outputs to ensure they meet the assessment requirements. Because AI models such as ChatGPT do not have a transparent search function and it is often unclear where information comes from, searching for literature via academic databases such as Web of Science is strongly recommended. For other types of information, we also recommend using authentic and verifiable sources. For data analyses, AI may assist with calculations, visualization, or code suggestions, but it cannot replace your own interpretation, verification, or final analysis of results. Within the assignment, you may use AI tools for:

- Editing
  - Check narration or subtitles in videos/audio for grammar, clarity, and flow.
  - Review instructions, labels, or UI text in prototypes for clarity in the prototypes.
  - Check code comments, variable names, or textual descriptions in scripts, programming, or visualizations for readability and overall coherence.
- Inspiration and brainstorming: AI may be used to explore ideas. Verify all AI-generated information with reliable sources and reference these in your assignment. Incorrect, fabricated sources or missing references will not be tolerated.
- AI can support programming by suggesting solutions, assist with data analysis by performing calculations and visualizing data, and aid audio transcription by converting speech to text. Verify and edit all AI outputs to ensure accuracy, coherence, and alignment with assignment requirements.

It is not allowed to use AI to generate final products. It is a support tool. All content must be created, evaluated, verified, and edited by you. Any use of AI must be limited to assisting, not replacing, your own work.

- Transparency: Clearly document for each tool what you used it for in the technology statement. You must use the **Technology statement student(s)** (see text below). Incomplete transparency or presenting AI output as your own work is not permitted.
- Data confidentiality: Never enter sensitive or confidential data into AI platforms, as these may not comply with Tilburg University's or external organizations' privacy guidelines.

Improper or non-transparent use of AI within the assignment will lead to sanctions and may result in your assignment being declared invalid by the Examination Board.

**Technology statement student(s)**

For the assignment, you must add the technology statement, using the text below. Replace the text in capital letters with the requested information.

During the preparation of this work, I/We used [**NAME TOOL / SERVICE / VERSION OF AI TOOL**] in order to [**REASON**]. The following parts of the assignment were affected/generated by AI tool usage: [**INTRODUCTION / METHODS / xxx, DISCUSSION**]. After using this tool/service, [**NAME STUDENT(S)**] evaluated the validity of the tool's outputs, including the sources that generative AI tools have used, and edited the content as needed. As a consequence, [**NAME STUDENT (S)**] take(s) full responsibility for the content of their work.