Open | DSE

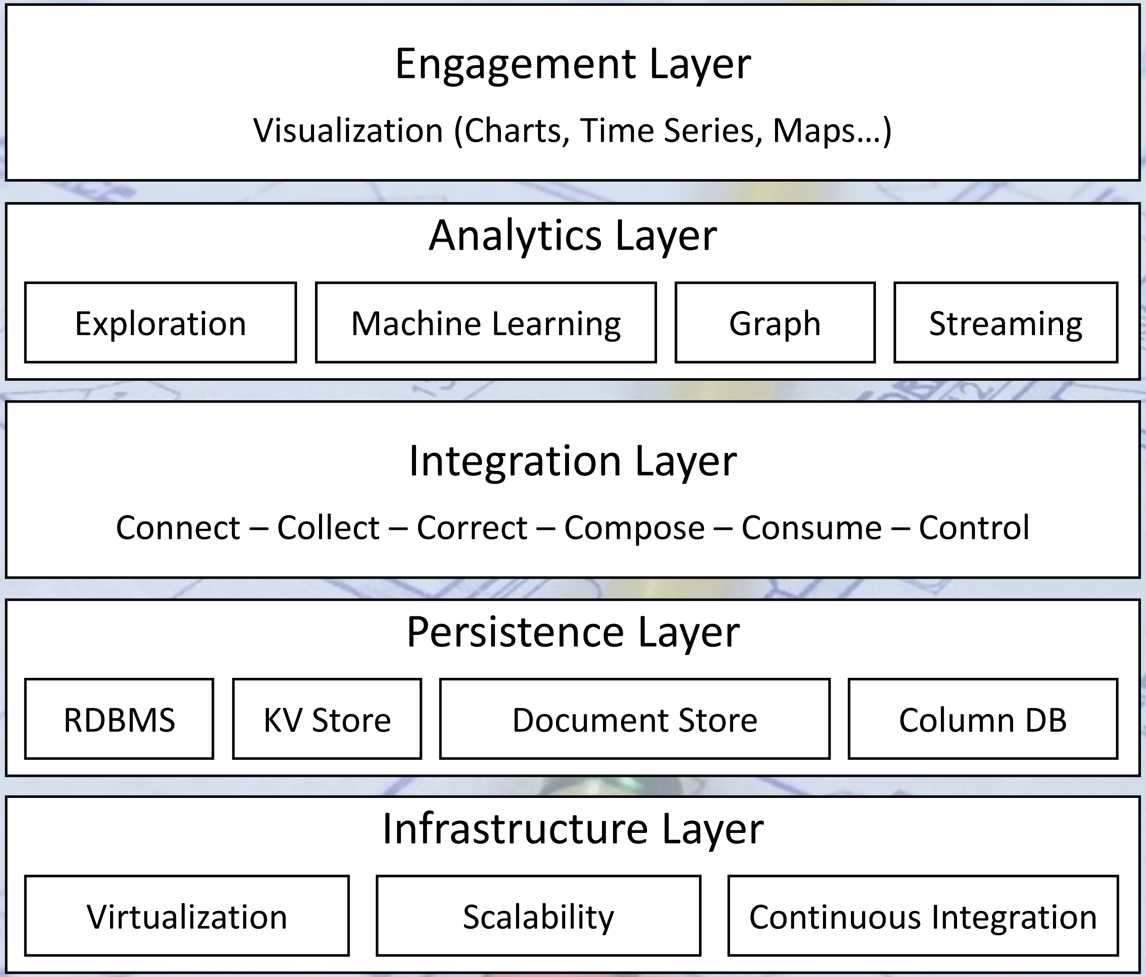
Architecting Big Data Solutions with Apache Spark
Lecture 2: Build Batch Applications

Ekhtiar Syed

Data Engineer/Scientist

Signify Research (formerly known as Philips Lighting)

DIA Architecture (Recap)



Engagement Layer

Visualization (Charts, Time Series, Maps…)

Analytics Layer

Exploration | Machine Learning | Graph | Streaming

Integration Layer

Connect – Collect – Correct – Compose – Consume – Control

Persistence Layer

RDBMS | KV Store | Document Store | Column DB

Infrastructure Layer

Virtualization | Scalability | Continuous Integration

Source: Spark for Python Developers – Amit Nandi

The engagement layer interacts with the end user and provides dashboards, interactive visualizations, and alerts.

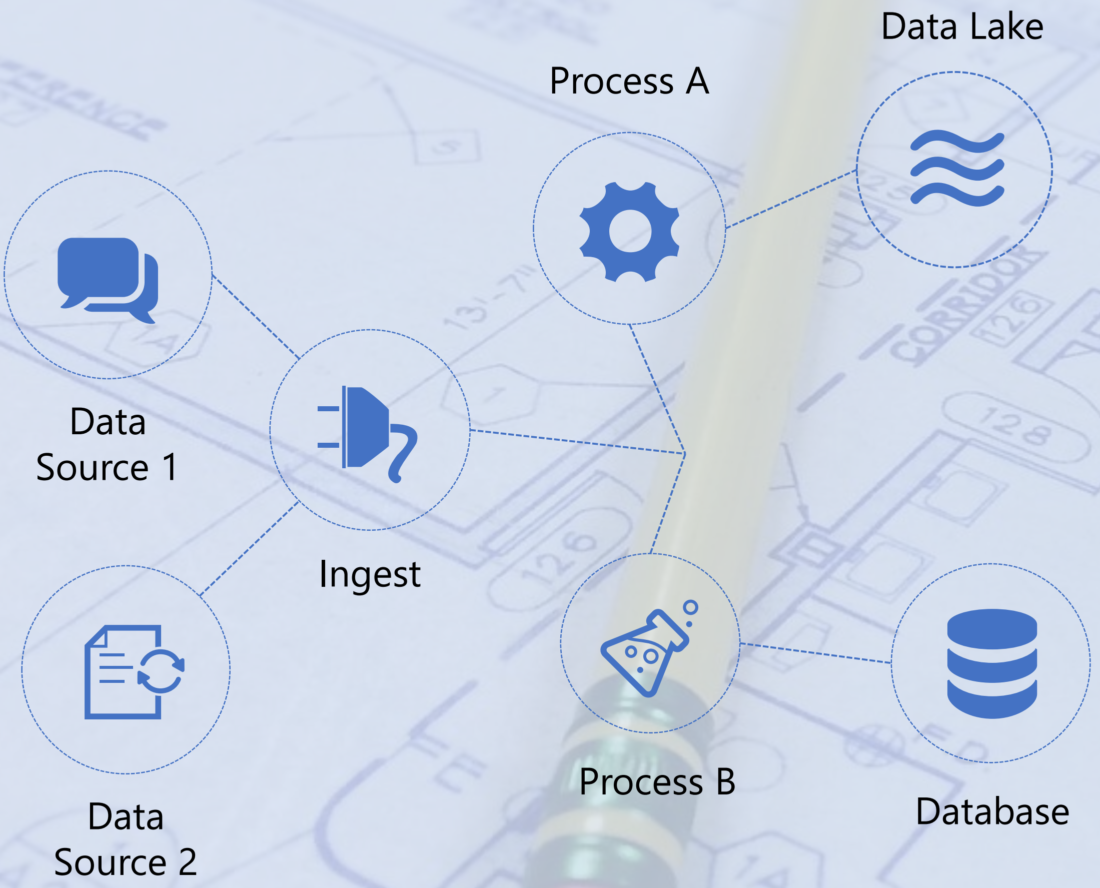
The analytics layer is where Spark processes data with the various models, algorithms, and machine learning pipelines in order to derive insights.

The integration layer focuses on data acquisition, transformation, quality, persistence, consumption, and governance. It is driven by the following five Cs: *connect*, *collect*, *correct*, *compose*, and *consume*.

The persistence layer manages the various repositories in accordance with data needs and shapes.

The infrastructure layer is primarily concerned with virtualization, scalability, and continuous integration.

What is a Data Pipeline?

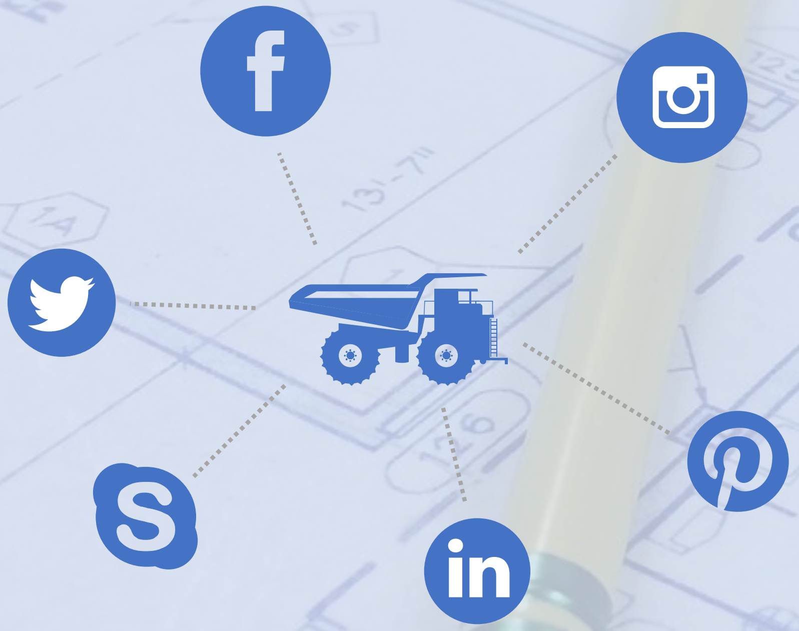


Traditionally, a pipeline is a collection of data processing tasks connected in a series, where the output of one task is the input of the next task. [1]

Data pipelines are a major part of DIA. Data pipelines in real-world settings typically consist of multiple tasks leveraging different technologies to meet required design goals or considerations.

[1] in Encyclopedia Of Information Technology, New-Delhi, Atlantic Publishers & Dist, 2007 , p. 382.

Data Pipeline in Companies



Netflix has a **data pipeline** to process **1.3 petabyte** of data per day to enable features like movie recommendation [1].

Facebook's real time data pipeline powers use cases like insights for Facebook page and analytics for mobile applications [2].

Twitter has a data pipeline to use **deep learning** at scale and show the **best Tweets** for your timeline [3].

[1] https://medium.com/netflix-techblog/evolution-of-the-netflix-data-pipeline-da246ca36905 [2] https://research.fb.com/publications/realtime-data-processing-at-facebook/ [3] https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines.html

Types of Data Pipelines

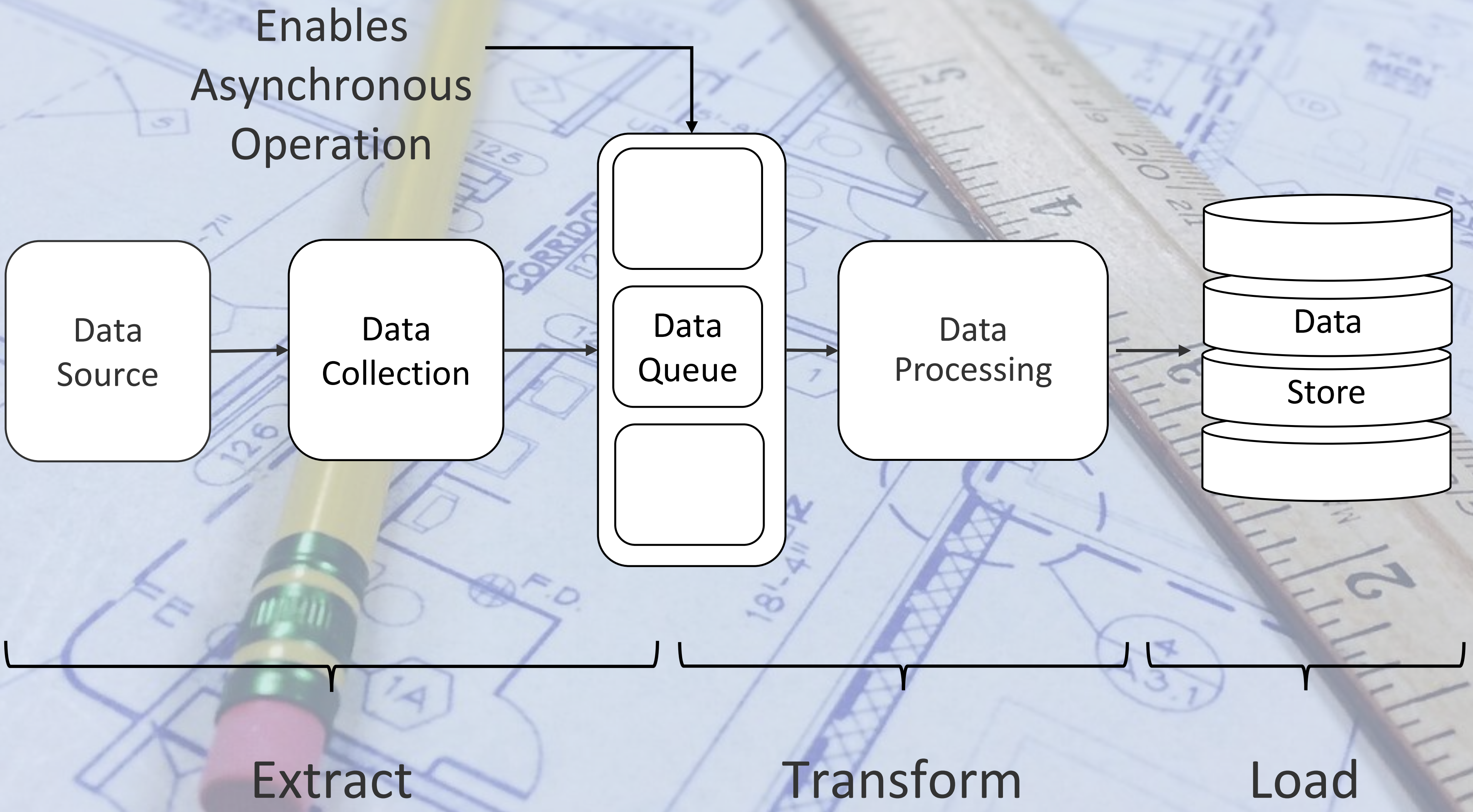
Batch / ETL Data Pipelines

Streaming Data Pipelines

Open | DSE

Architecting and Implementing Batch Oriented Data Pipelines

Reference Architecture: Batch Oriented Data Pipelines



Enables Asynchronous Operation

Data Source

Data Collection

Data Queue

Data Processing

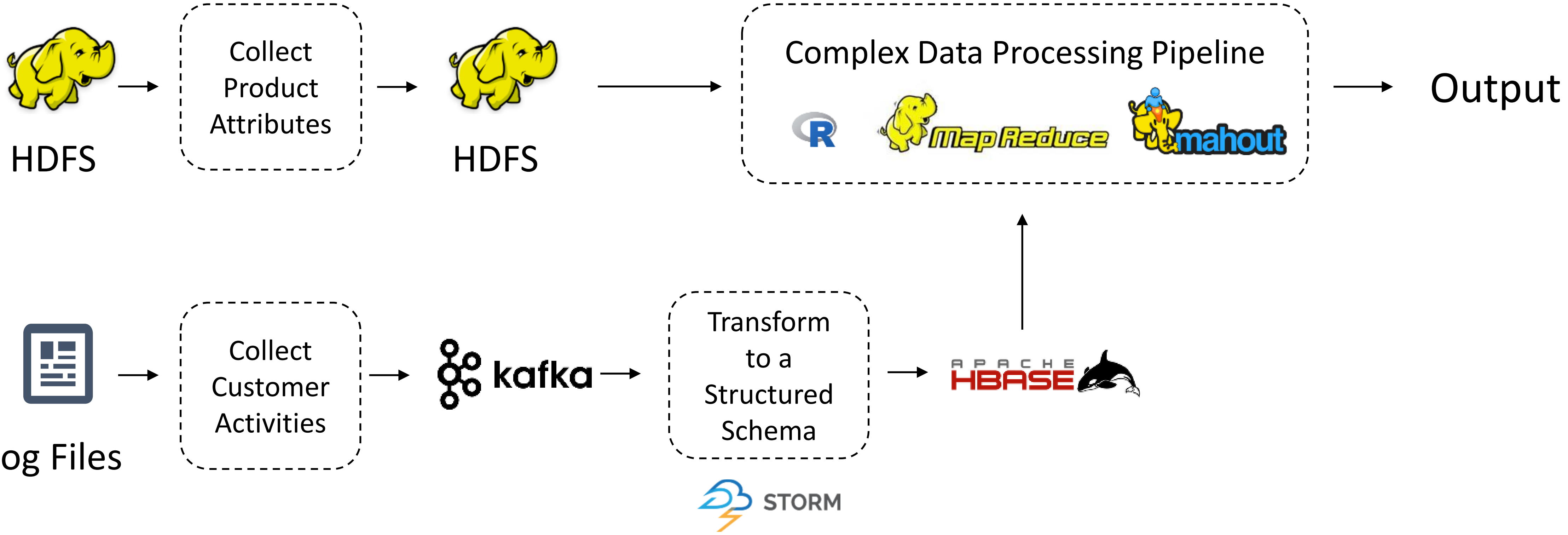
Data Store

Extract

Transform

Load

Groupon: CRM Data Gathering and Mining Pipelines



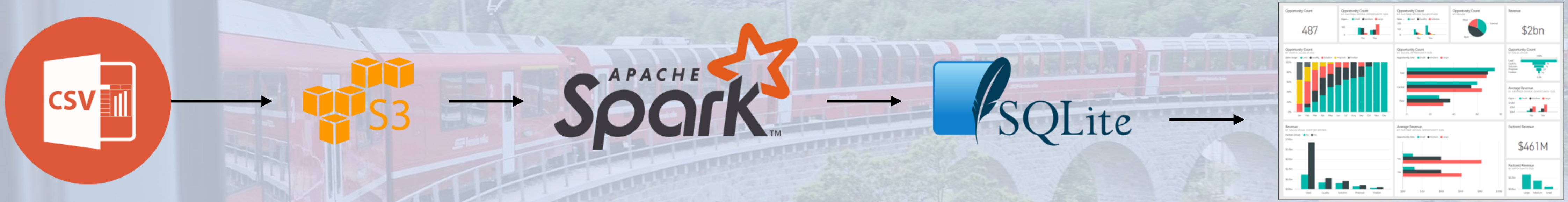
HDFS → Collect Product Attributes → HDFS → Complex Data Processing Pipeline → Output

Log Files → Collect Customer Activities → kafka → Transform to a Structured Schema → HBASE → Complex Data Processing Pipeline

V. D. a. N. P. Kang Li, "Big Data Gathering and Mining Pipeline for CRM using Open-source," in *IEEE International Conference on Big Data*, Dalian, 2015.

Our Project

Link: https://opentransportdata.swiss/en/dataset/istdaten



SBB CFF FFS

Let’s Start The Practical Part!