GENERAL

* Data model: conceptual, logical, physical
* Pipeline

Diagram

Description automatically generated

* Type of processing (batch) – eventual type of scaling if data were bigger (recall that relational DBs are hard to scale horizontally -fuck)
* Analyse the 5Vs of big data
  + Volume
  + Velocity - from static data to data streams
  + Variety - mixing and matching different types of data
  + Veracity - quality
  + Value – extracting knowledge from data
* Do we want to do a Data Quadrant for the report?
* Type of license and comply with license requirements (see and expand file in Git)

STORING BIGDATA

* DBs are made for fast info retrieval, not for storing (is this the case in our case?)
* NoSQL vs SQL: why we chose relational DB
* Distributed systems: cap theorem ?Do we need that?

ACCESSING/RETRIEVING BIGDATA

* We use SQL

INGESTING BIGDATA

* Challenges of data ingestion
  + Volume: need to be able to ‘digest’ fat data
  + Velocity: data may come in very fast
  + Variety: data may come in a variety of formats and models, and they can change often over time
* Data ingestion parameters (to consider for building an efficient ingestion pipeline)
  + Data Velocity – the speed at which data flows in from different sources
  + Data Size – how big is the data today? How big will it be tomorrow?
  + Data Frequency - batch or stream (real-time)
  + Data Format (Structured, Semi-Structured, Unstructured)
* Pubs/hubs: do we need them?
* Apache Spark: do we need it?