Data Lake Ingestion checklist

By Dmytro Dragan

softserve

BIG DATA LEAD ENGINEER

DMYTRO DRAGAN

SUMMARY OF QUALIFICATIONS

Dmytro Dragan has 9 years of experience in Software Design and Development, where he converts client`s ‘want to’ in production-ready solutions.

Last 5 years his focus lies on building data pipelines on open-source frameworks and tools.

He’s also worked on quite disparate things such as real-time anti-fraud system, phone gpu image processing, machine learning modeling.

TECHNOLOGIES

- RDBMS: Oracle, MS SQL Server, DB2
- NoSQL: Redis, Cassandra
- Hadoop stack: Hive, Impala, HBase, Spark
- AWS stack: AWS Glue, Lambda, Athena, EMR, DynamoDB, Redshift
- GCP stack: Dataflow, Cloud Function, Dataproc, Composer, BigTable, BigQuery
- Streaming Frameworks: Flink, Kafka-Streams, Flume, Storm
- ML: Jupyter, TensorFlow, H2O, R

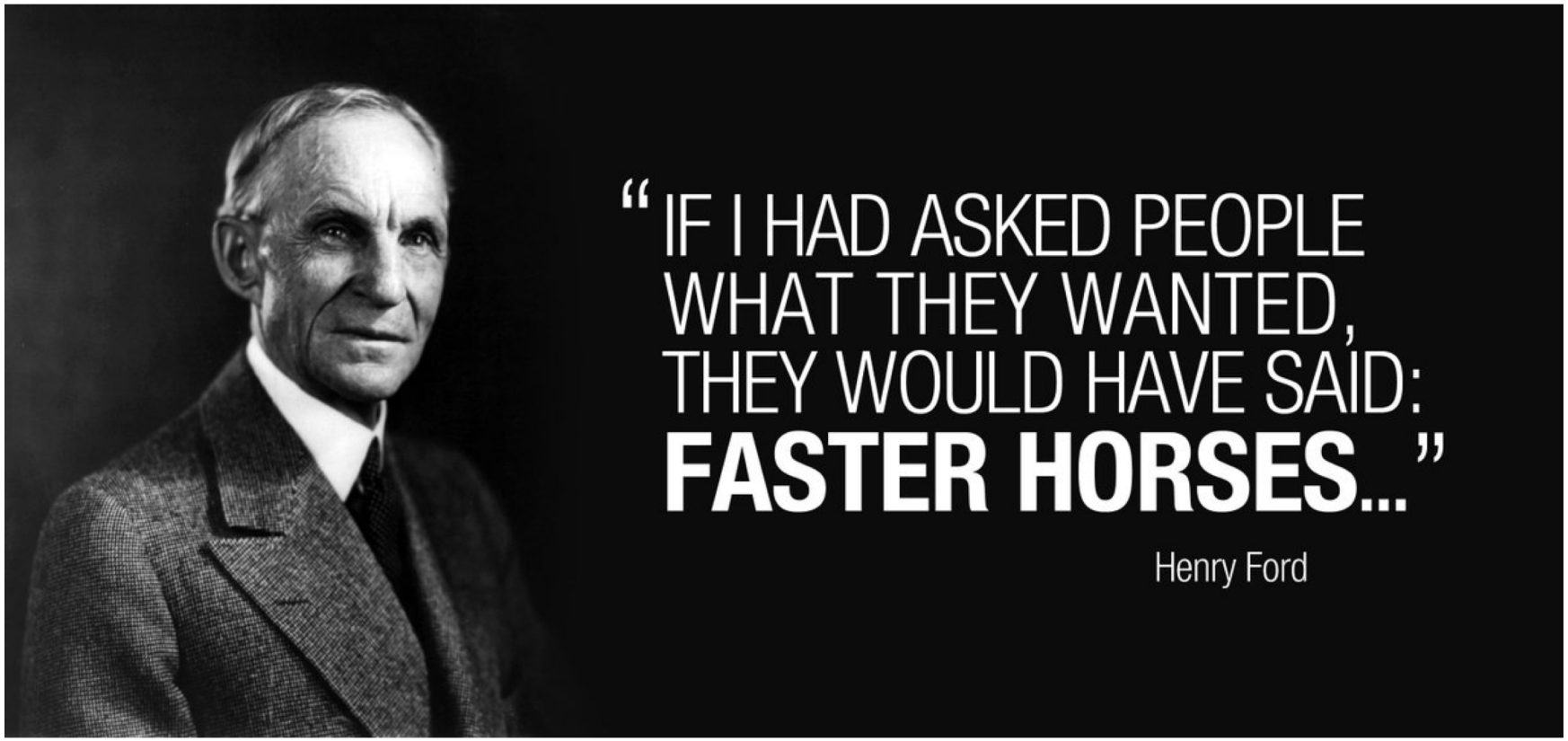
softserve

WHAT IS OUR GOAL?

WHAT IS OUR GOAL?

- To understand what clients want
- To give clients what they want

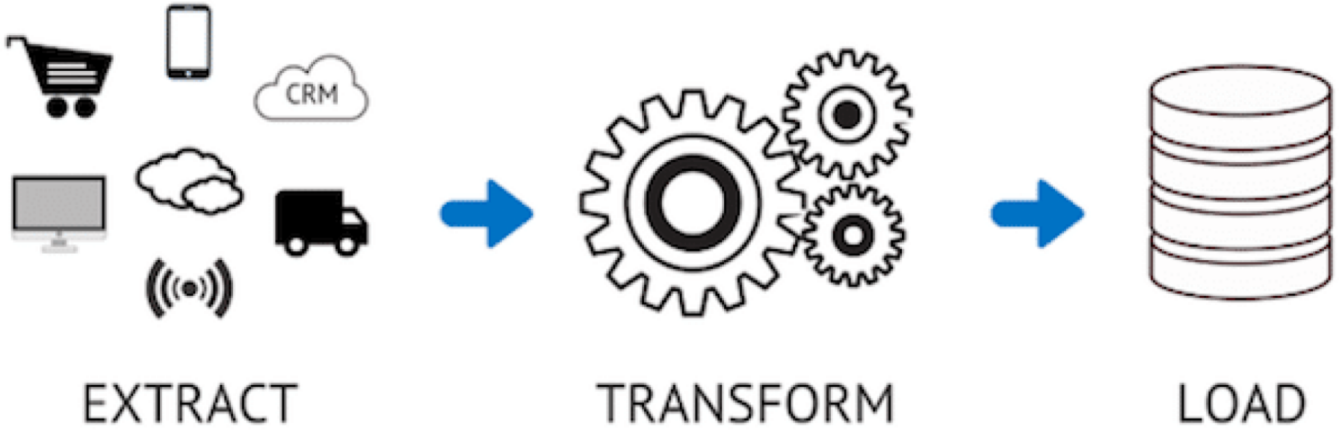
I like this example



WHAT IS OUR GOAL?

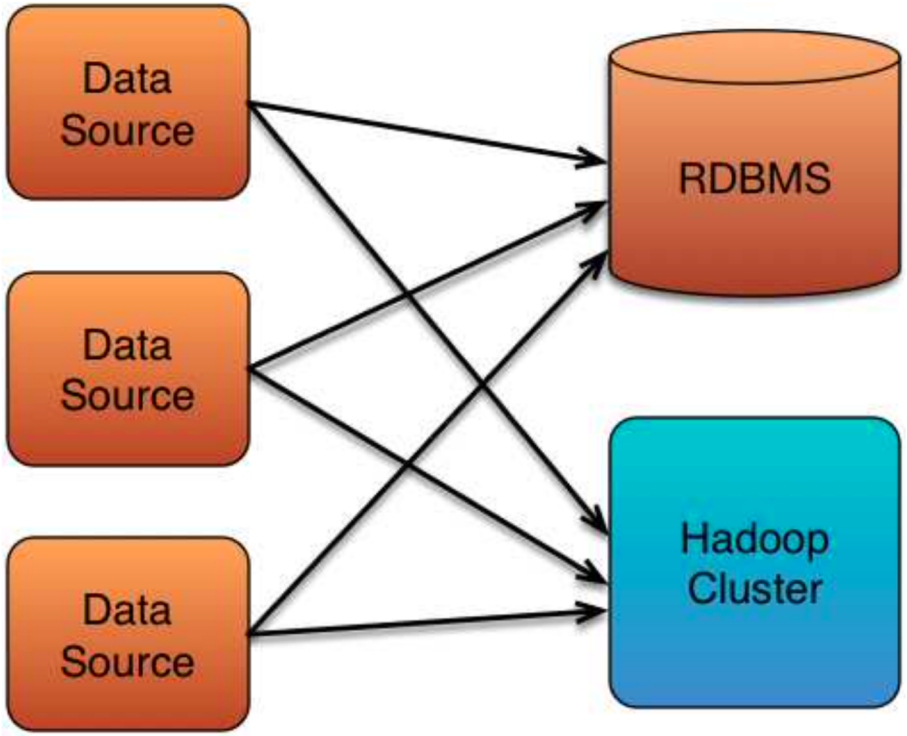
- To understand what clients want
- ~~To give clients what they want~~
- To understand what clients need
- Show them both
- Help client to choose

Ingestion is simple... right?



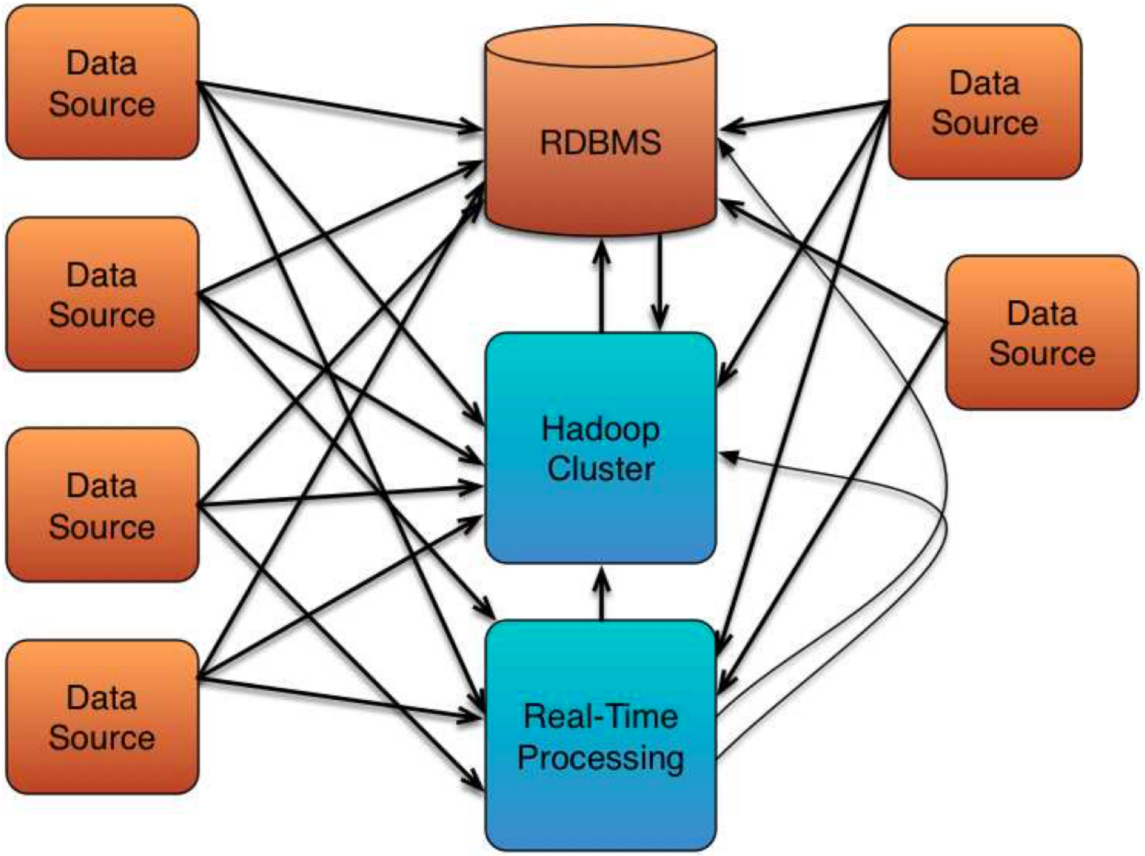
Ingestion is simple... right?

It is (relatively) easy to connect just a few systems together



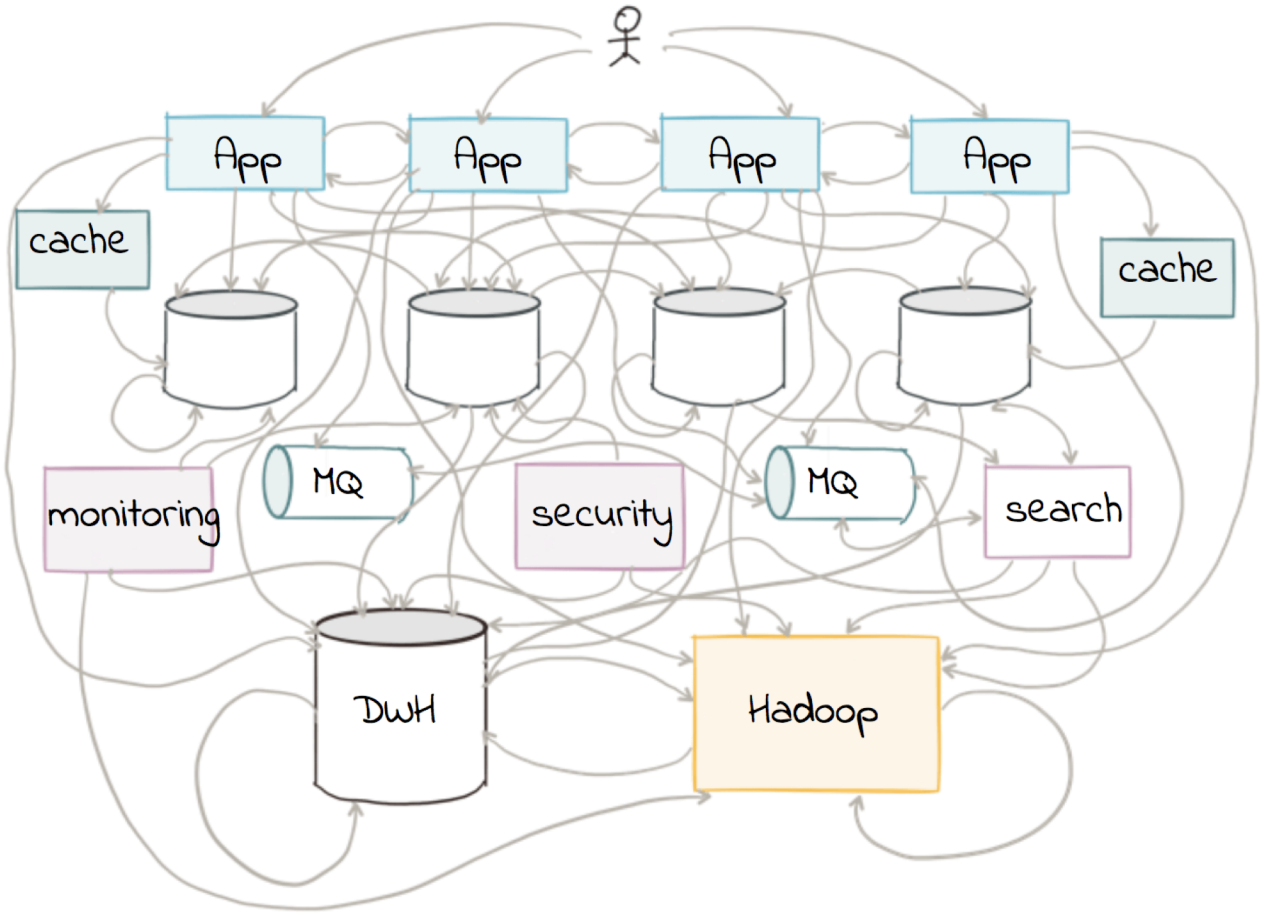
Ingestion is simple... right? (Not really)

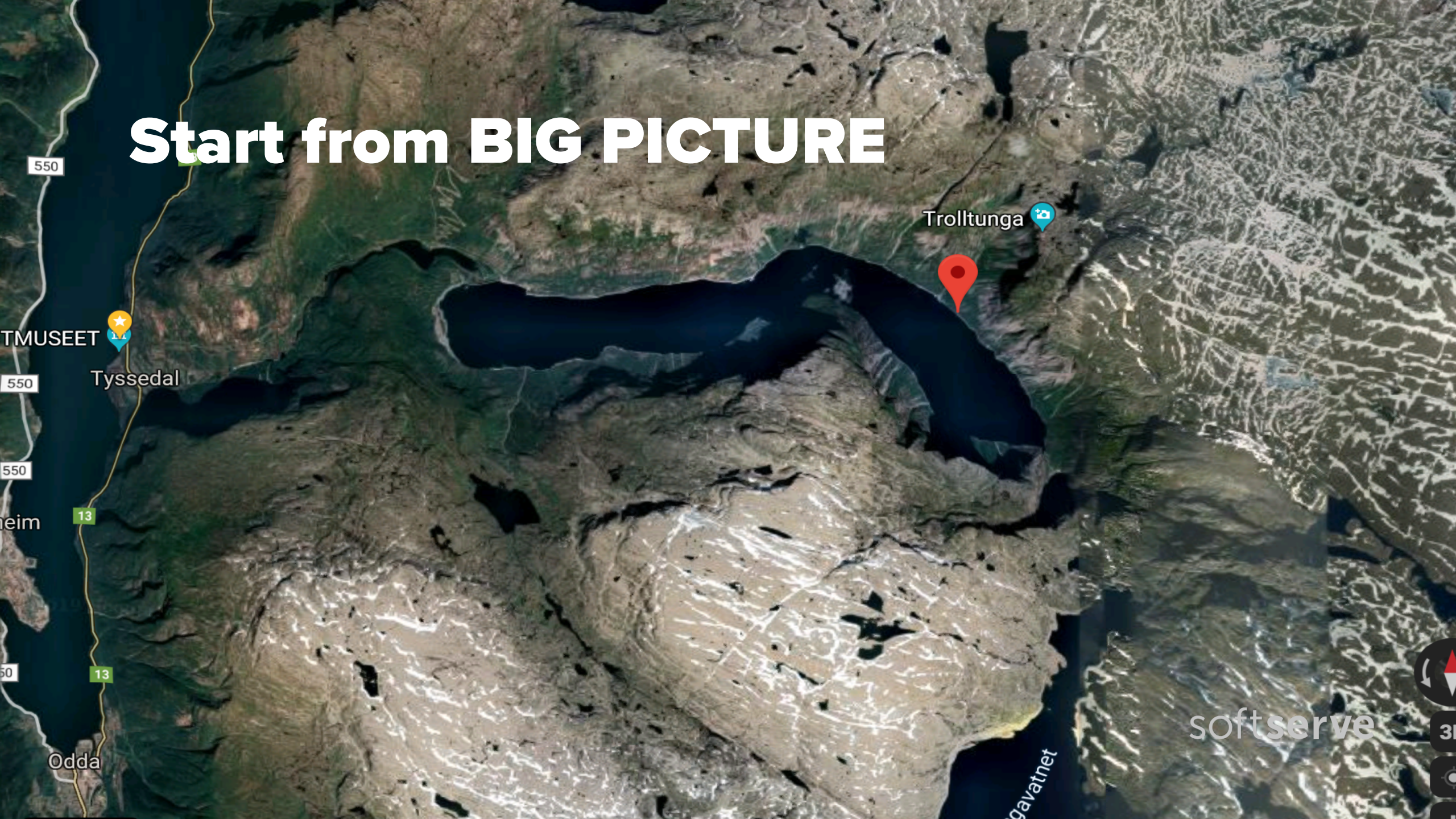
As we add more systems, complexity increases dramatically



Ingestion is simple... right? (Nope)

...until eventually things become unmanageable




Start from BIG PICTURE
550
Trolltunga
TMUSEET
Tyssedal
550
550
13
13
Odda
softserve

Start from BIG PICTURE

- The Client motivation
  - Ask ”why”
  - Find out what are the driving requirements/objectives
- 97: Seek the value in requested capabilities
- 97: Diminish accidental complexity

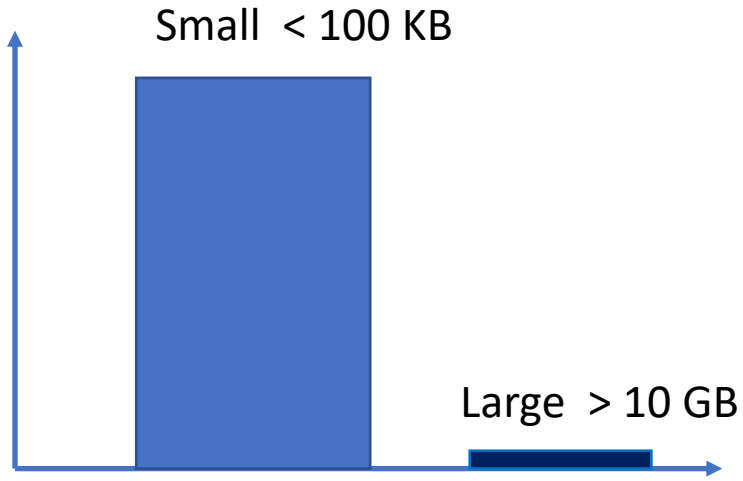
https://learning.oreilly.com/library/view/97-things-every/9780596800611/

Short Story #0 – Spark is Gold Hammer

”There will be a large files, so we definitely need Spark” Customer

Short Story #0 – Spark is Gold Hammer

”There will be a large files, so we definitely need Spark” Customer

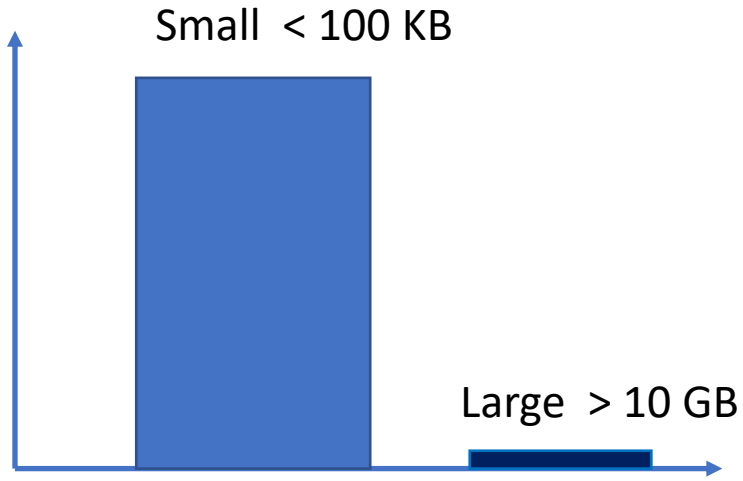


1000 data feeds are small (less 100 KB):

- Spark cluster start time will be more time than processing
- Bad cluster utilization

Short Story #0 – Spark is Gold Hammer

”There will be a large files, so we definitely need Spark” Customer

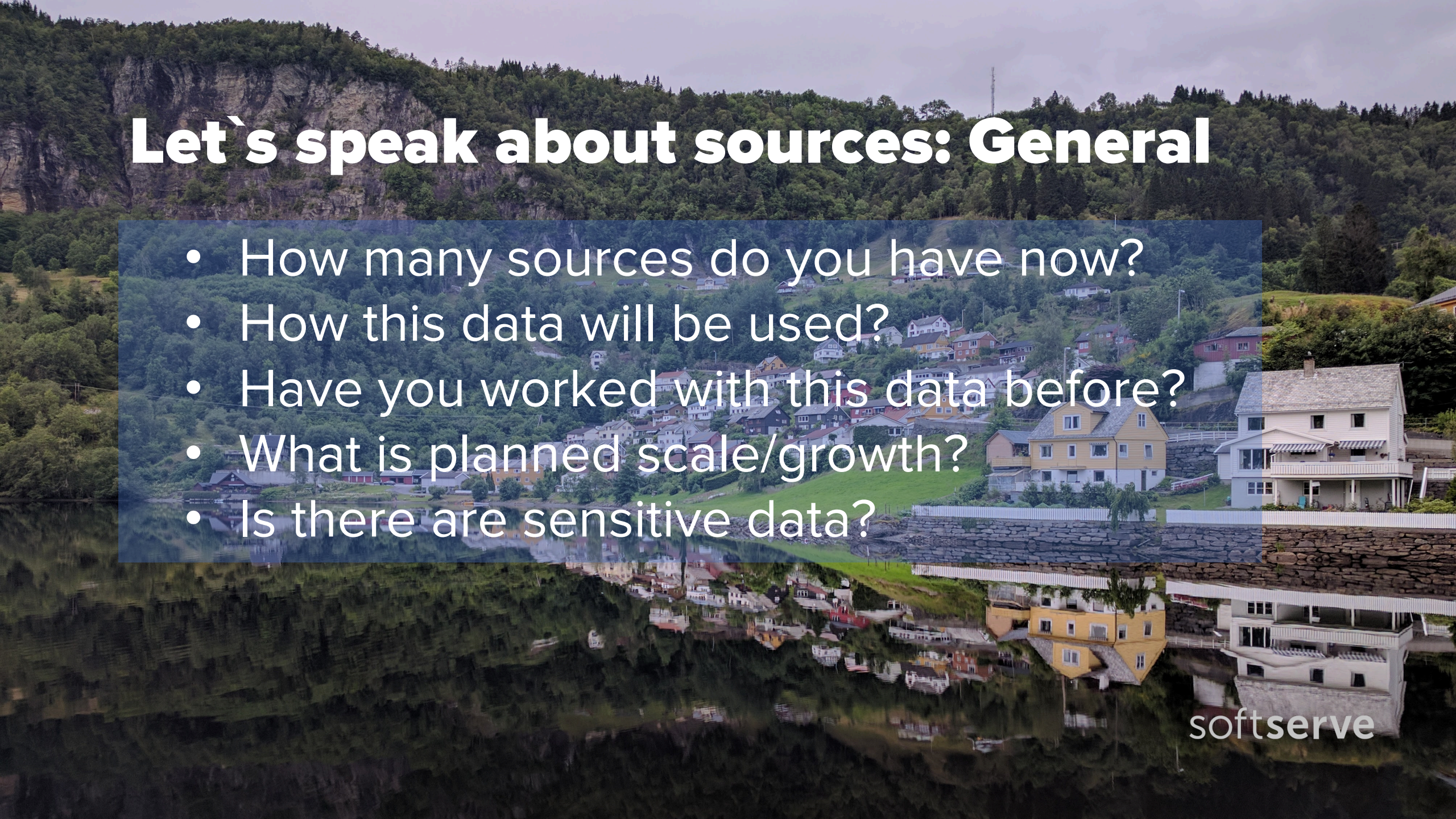


1000 data feeds are small (less 100 KB):

- Use Spark for only for heavy jobs
- Use light-weight solution for others

Let`s speak about sources

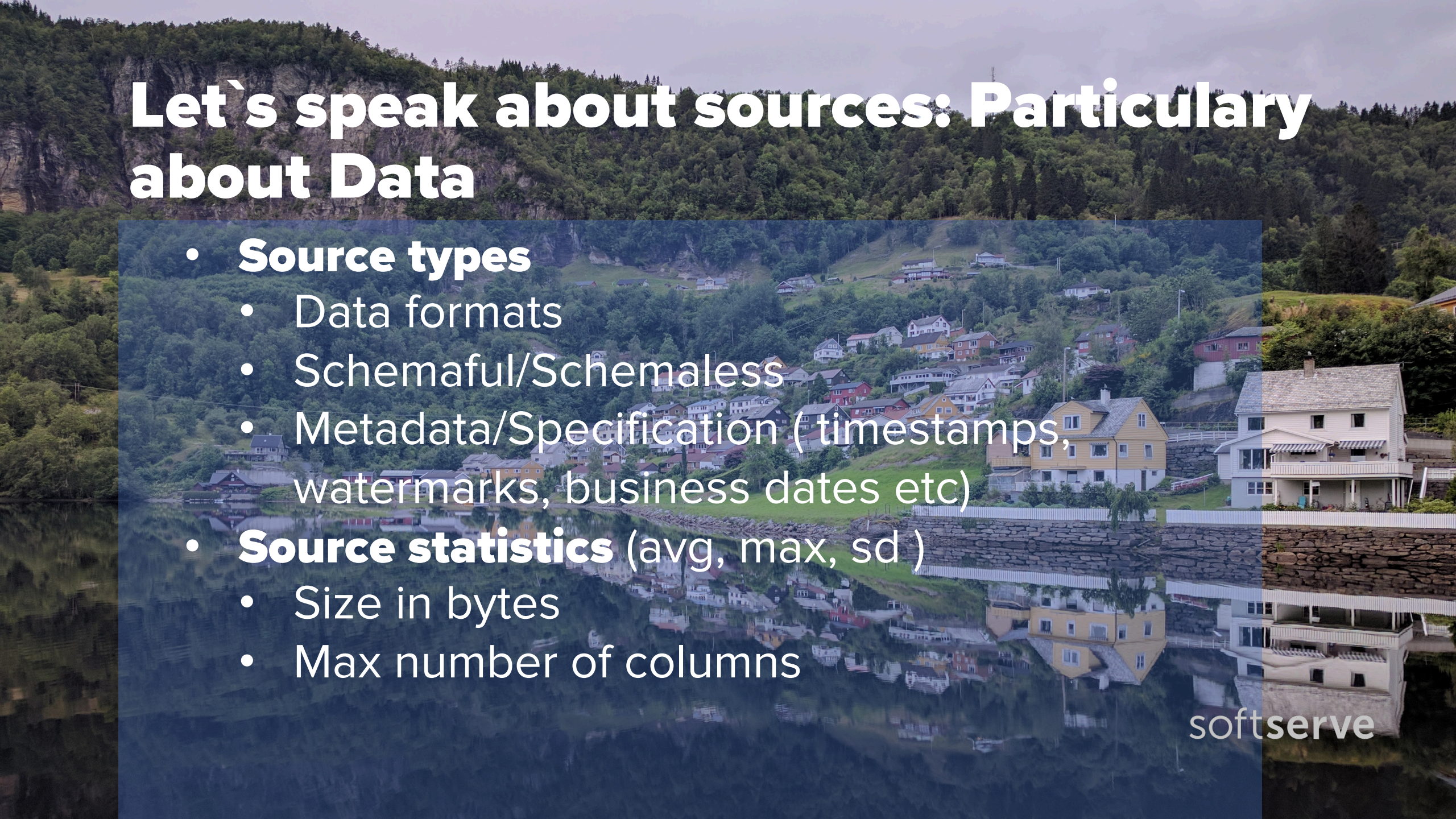
softserve



Let`s speak about sources: General

- How many sources do you have now?
- How this data will be used?
- Have you worked with this data before?
- What is planned scale/growth?
- Is there are sensitive data?

softserve



Let`s speak about sources: Particulary about Data

- **Source types**
  - Data formats
  - Schemaful/Schemaless
  - Metadata/Specification (timestamps, watermarks, business dates etc)
- **Source statistics** (avg, max, sd)
  - Size in bytes
  - Max number of columns

softserve

Let`s speak about sources: Particulary about Source system

- **Frequency**
  - time-based
  - trigger-based (event-based, One-time-delivery, Only God knows)
- How long dataset will be available?
- Load on source system:
  - Number of parallel connectors
  - Direct connectors
  - Read window

softserve

Short Story #1 – Is it really time-based?

„We will provide file in last day of month at 1 a.m. CET” Customer

Short Story #1– Is it really time-based?

„We will provide file in last day of month at 1 a.m. CET,
if it will be weekend day, then in first working day.” Customer

Short Story #1 – Is it really time-based?

„We will provide file in last day of month at 1 a.m. CET,

if it will be weekend day, then in first working day.

If it will be bank holiday in Sweeden, then next working day.” Customer

Short Story #1 – Is it really time-based?

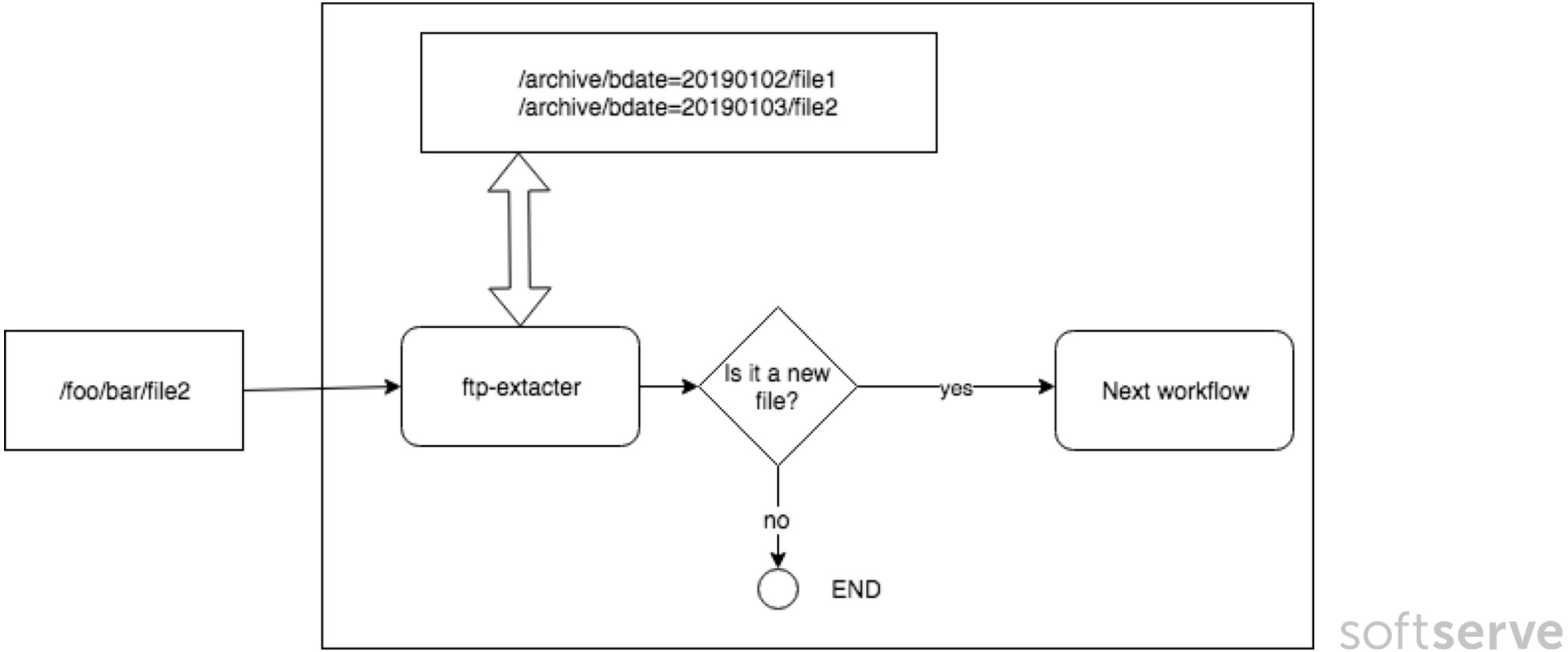
„We will provide file in last day of month at 1 a.m. CET,

if it will be weekend day, then in first working day.

If it will be bank holiday in Sweeden and files comes from Swedish MF system (you will need parse it first to get header with Sweeden code), then next working day.

If it will be bank holiday in Norway and files comes from Norway MF system (you will need parse it first to get header with Norway code), then next working day.” Customer

Short Story #1 – Is it really time-based?



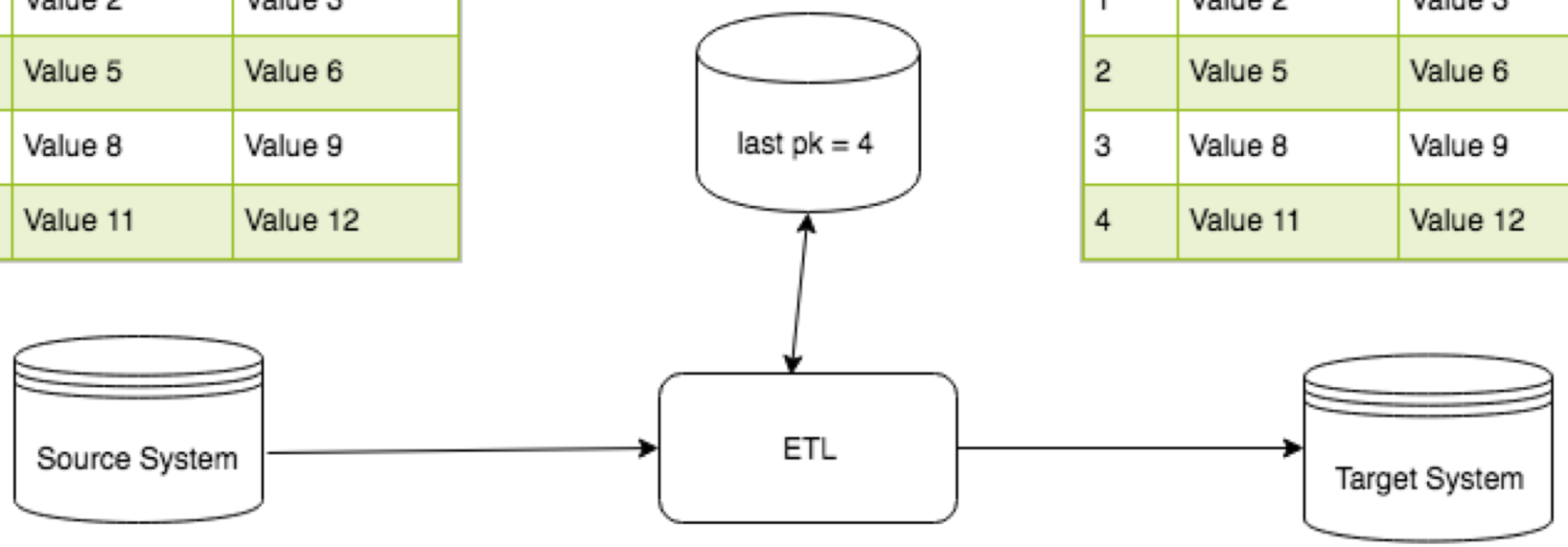
Let`s speak about sources: Data evolution

- Is input data append-only?
  - Full ingest
  - Incremental
  - Mix of both
- Schema evolution

Short Story #2 – It is a simple JBDC source, you already suport it

Incremental Slowly Changing Dimensions (SCDs) 2-6 Types

pk	Title 2	Title 3
5	Value 2	Value 3
6	Value 5	Value 6
7	Value 8	Value 9
8	Value 11	Value 12



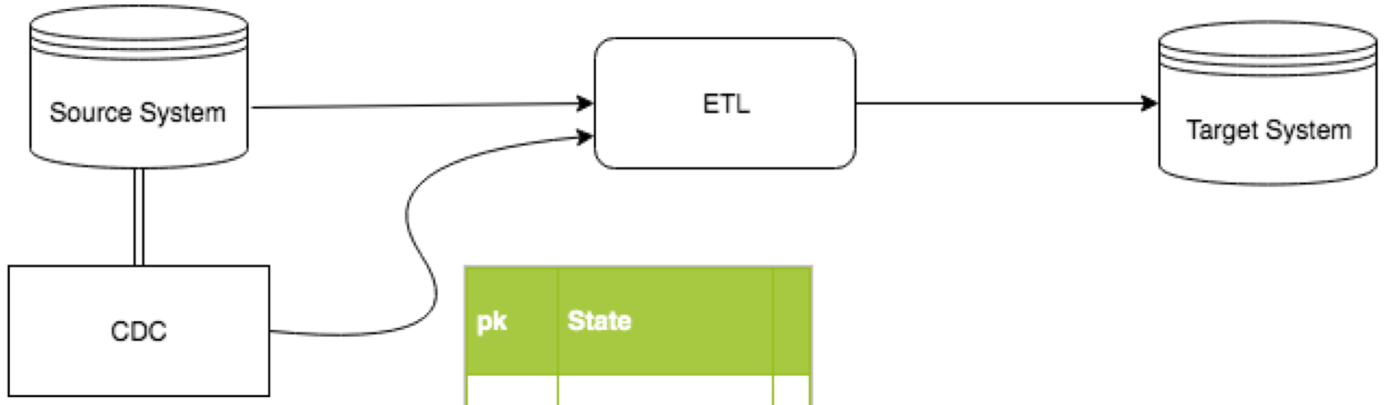
pk	Title 2	Title 3
1	Value 2	Value 3
2	Value 5	Value 6
3	Value 8	Value 9
4	Value 11	Value 12

Short Story #2 - It is a simple JBDC source, you already suport it

Incremental Snapshoting with Capture Data Change (CDC)

pk	Title 2	Title 3
1	Value 2	Value 3
2	Value 10	Value 66
3	Value 8	Value 9

pk	Title 2	Title 3
1	Value 2	Value 3
2	Value 5	Value 6
3	Value 8	Value 9
4	Value 11	Value 12



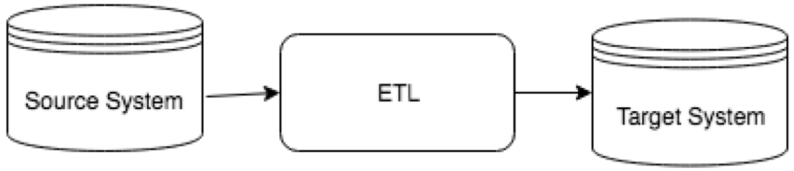
pk	State	
2	Updated	
4	Dropped	

softserve

Short Story #2 - It is a simple JBDC source, you already suport it

Incremental Snapshoting with Merge

pk	Title 2	Title 3
1	Value 2	Value 3
2	Value 10	Value 66
3	Value 8	Value 9



Snapshot 1

pk	Title 2	Title 3	Version
1	Value 2	Value 3	1
2	Value 5	Value 6	1
3	Value 8	Value 9	1
4	Value 11	Value 12	1

Snapshot 2

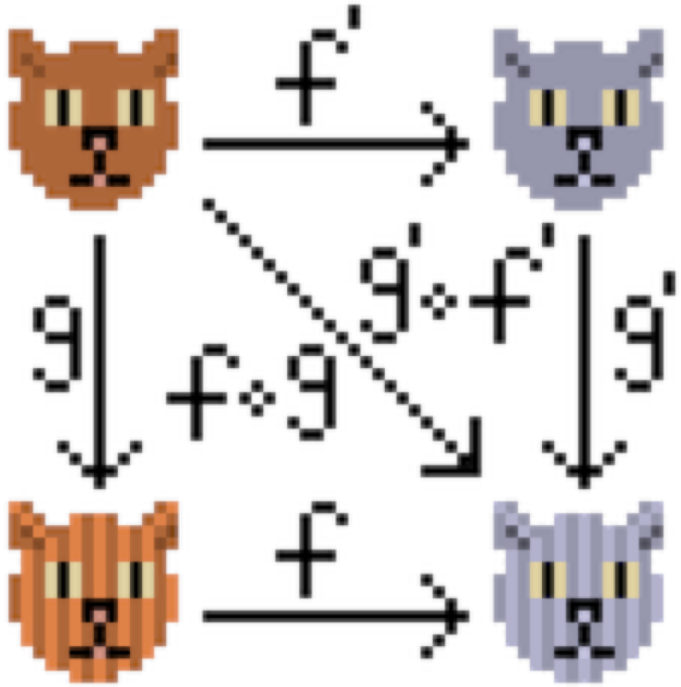
pk	Title 2	Title 3	Version
1	Value 2	Value 3	2
2	Value 10	Value 66	2
3	Value 8	Value 9	2

Final View

pk	Title 2	Title 3	Version
1	Value 2	Value 3	2
2	Value 10	Value 66	2
3	Value 8	Value 9	2
4	Value 11	Value 12	1

Transformations

- Variety
- Do you have dependencies ?
- State
- Fan-In/Fan-Out
- Field Mapping

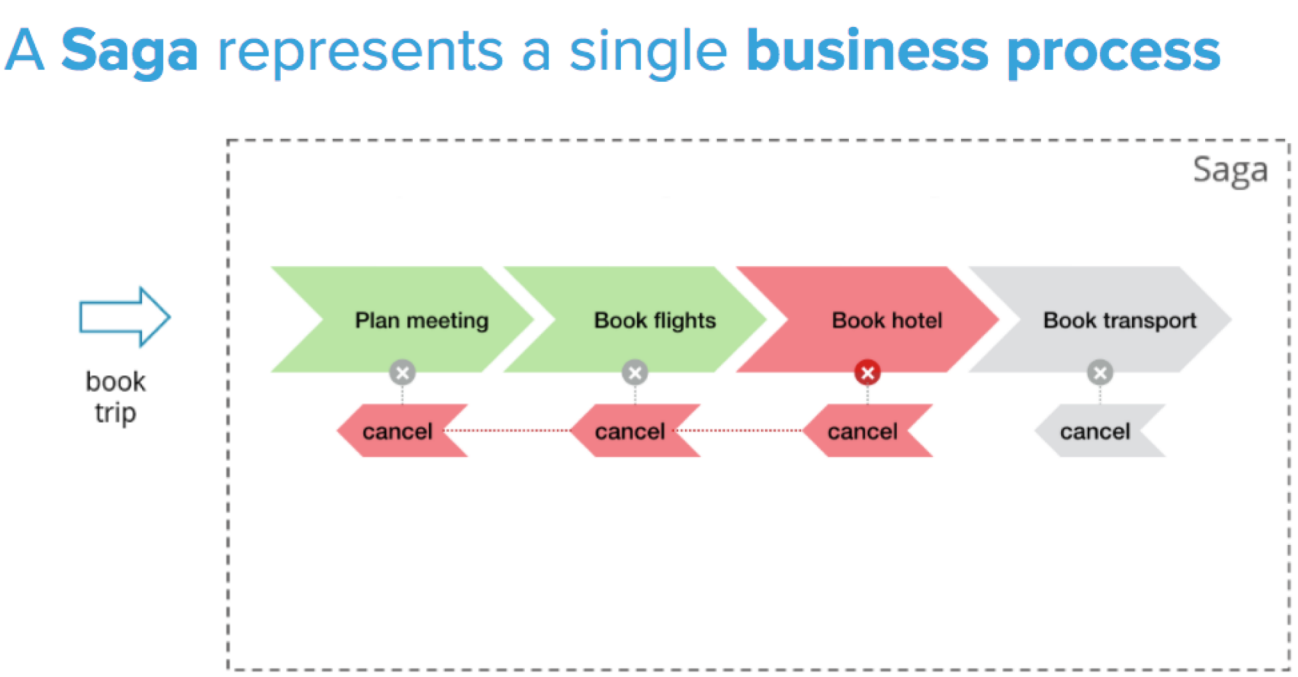


Short Story #3 – The Great Field Names

1. FILLER (duplicated 10 Times)
2. „THE SOME STRANGE FIELD”
3. *Døðr /* Колонна / 檐柱
4. 02FIELD (say ”HI” to AVRO)
5. …

Additional flows

- Exception flow
- Investigation flow



What about our target?

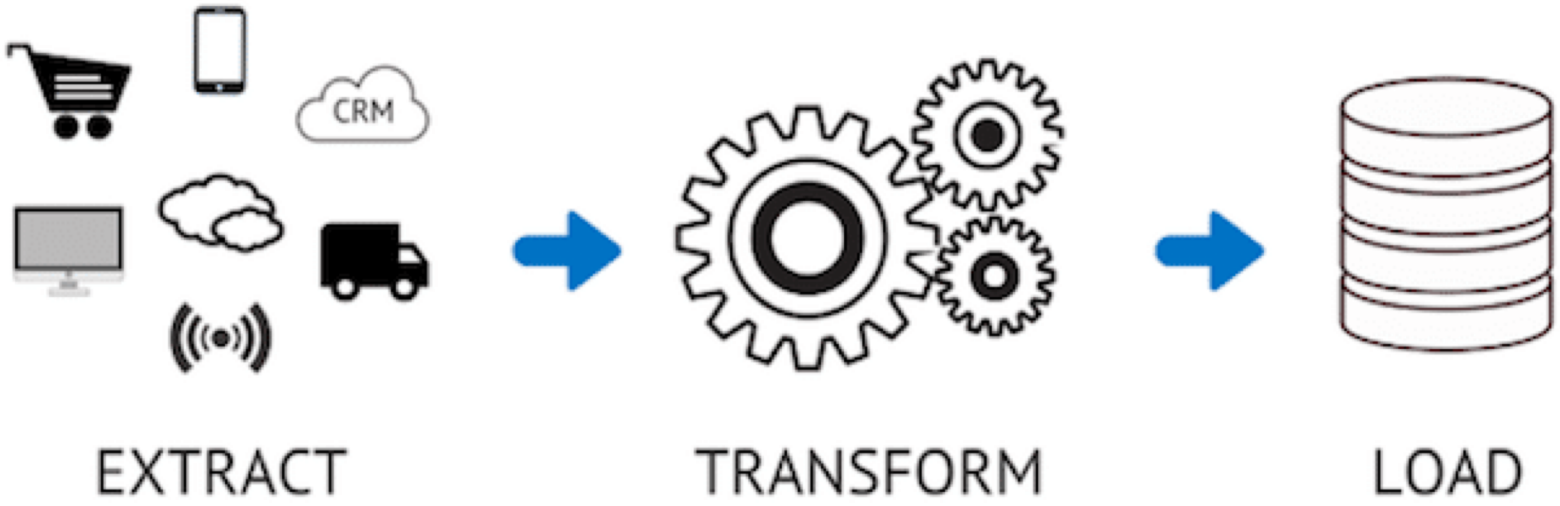
Now you are source system ☺

- **End-User Delivery Interfaces**
- Idempotent write
- SLA

Who else is your stakeholder?

- Data Governance / Data Quaility
- Data Lineage / Data Catalog
- Complience/Regulations
- Monitoring/ Log management

Any Questions?




Thank you
softserve