Visual Stethoscopes

Authors:
Juan Carlos Rosito Cuellar - j.rositocuellar@campus.unimib.it
Lida Amalia Follari - l.follari@campus.unimib.it

Data Management Final Project

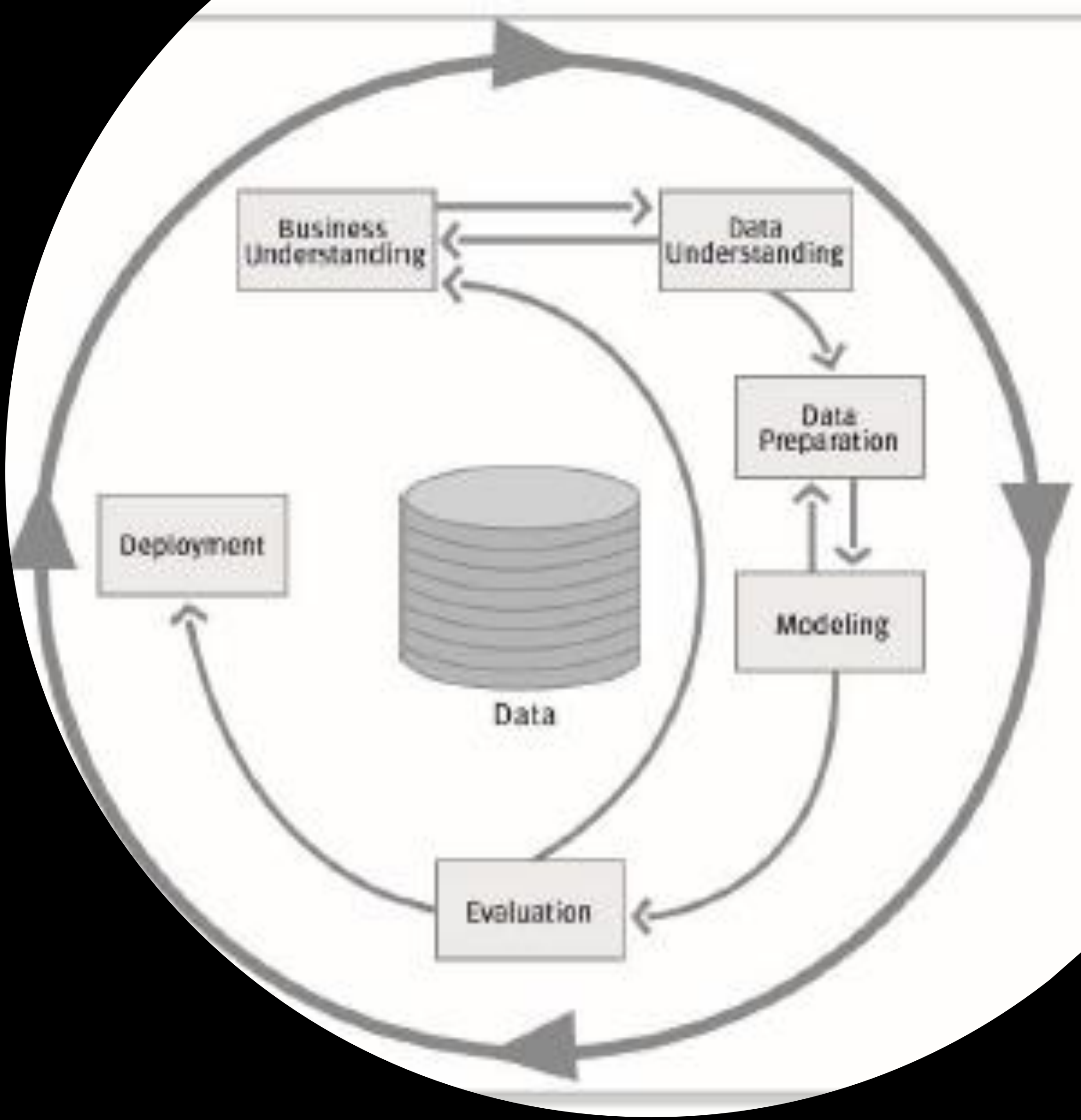
Prof. Maurino

Università Bicocca, Data Science, 2017-2018

Ciclo di Vita del Dato

Crisp-DM Reference Model

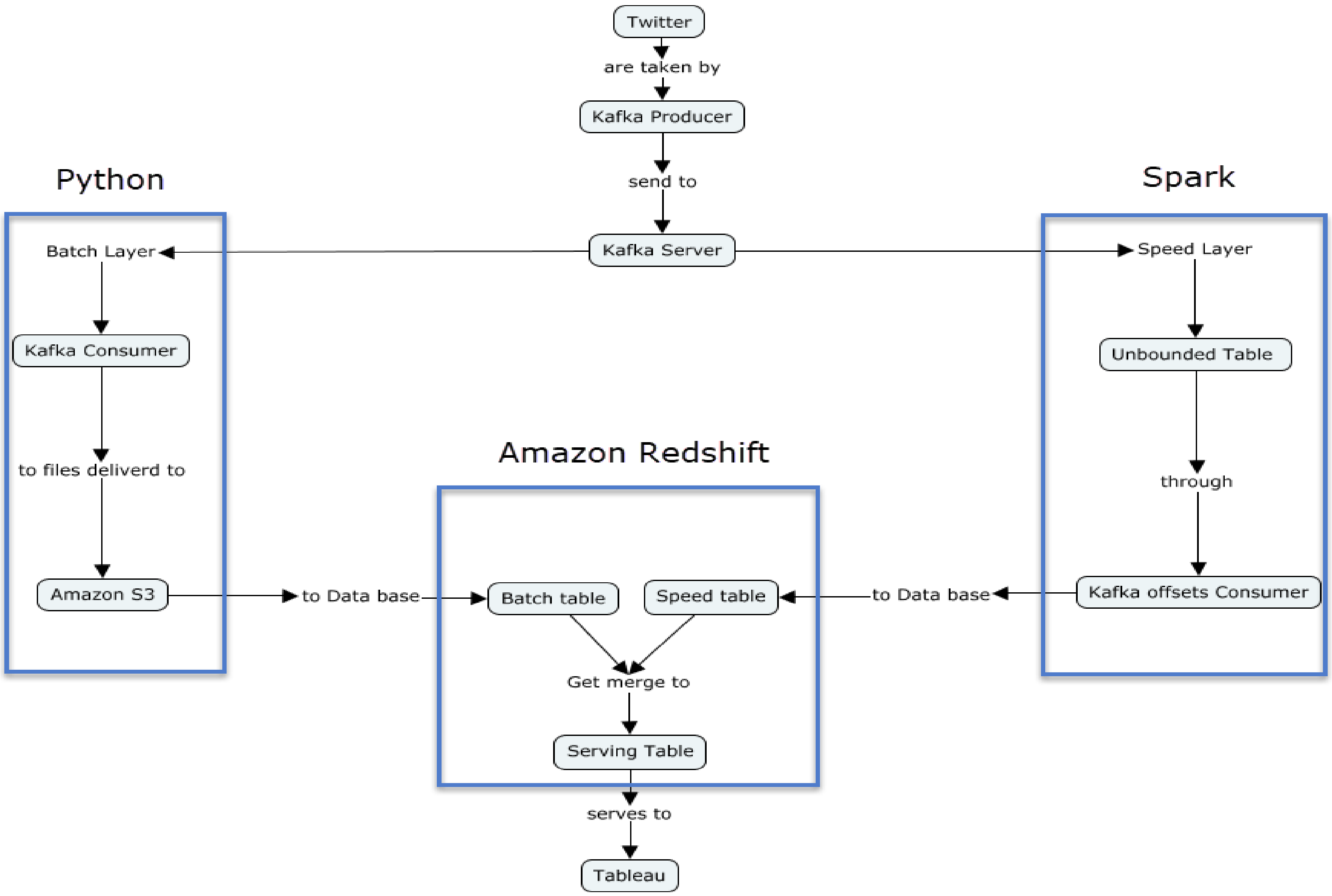
- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment



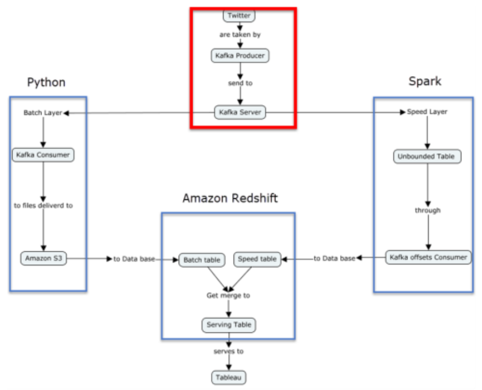
BUSINESS UNDERSTANDING: Visual Stethoscopes

- Listening to online conversations
- In real-time
- Discovering public opinion sensibility about some social indicators

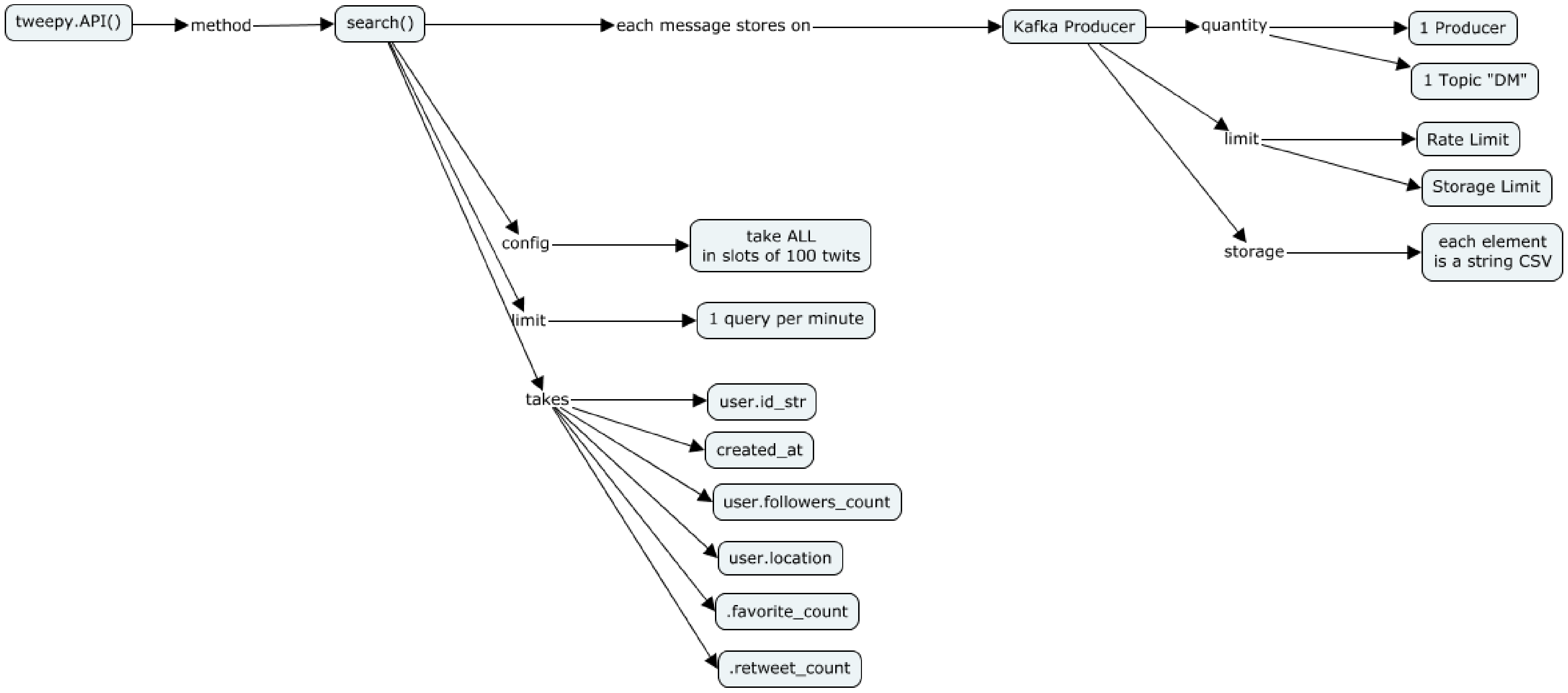
System Implementation
Lamda Architecture Pipeline

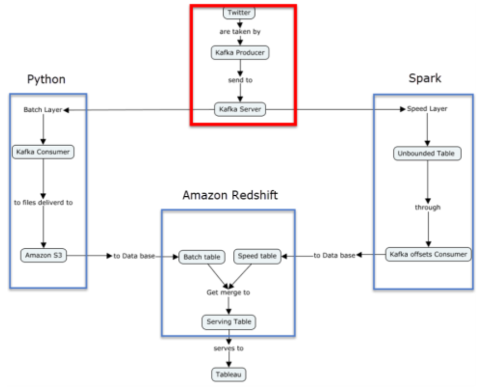


Data Ingestion



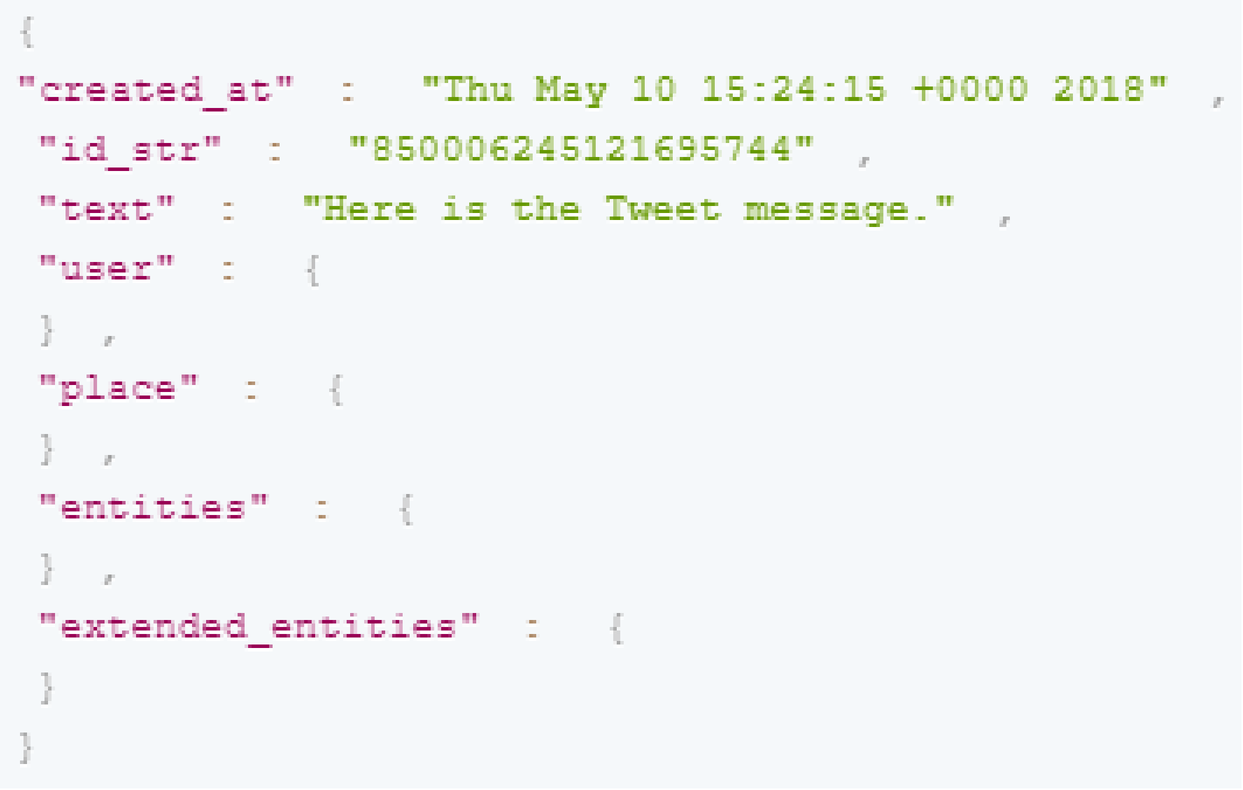
- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment





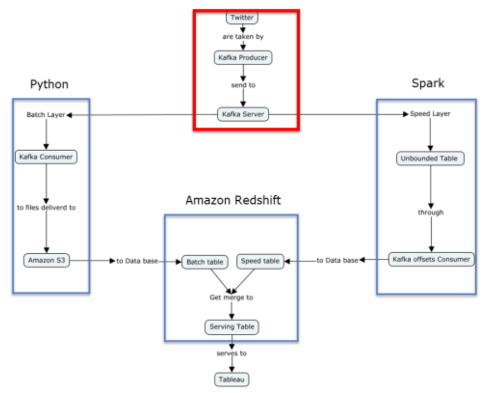
JSON Tweet

- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

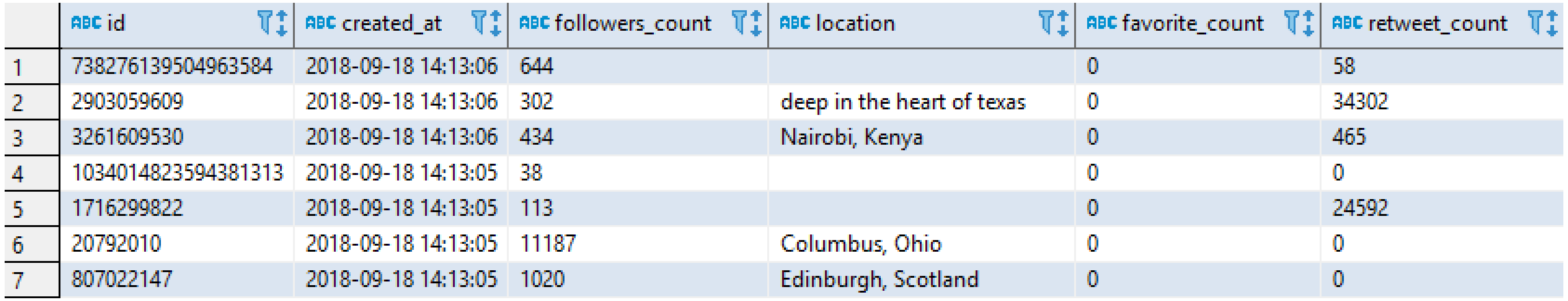


```
{
"created_at" :  "Thu May 10 15:24:15 +0000 2018" ,
 "id_str" :  "850006245121695744" ,
 "text" :  "Here is the Tweet message." ,
 "user" :  {
 } ,
 "place" :  {
 } ,
 "entities" :  {
 } ,
 "extended_entities" :  {
 }
}
```

Tweet Table

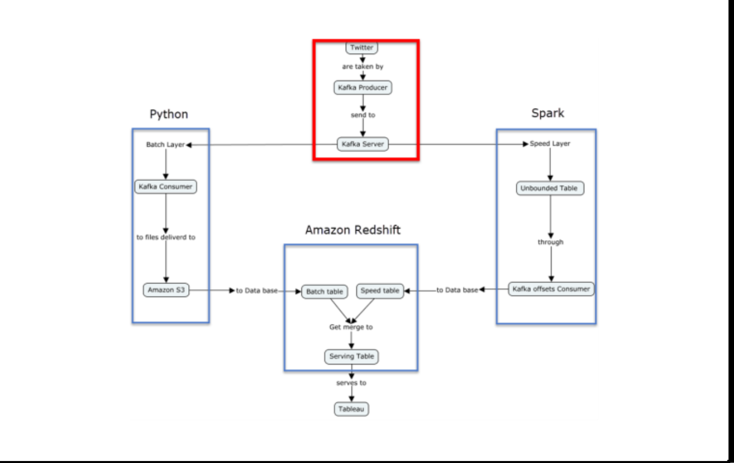


- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

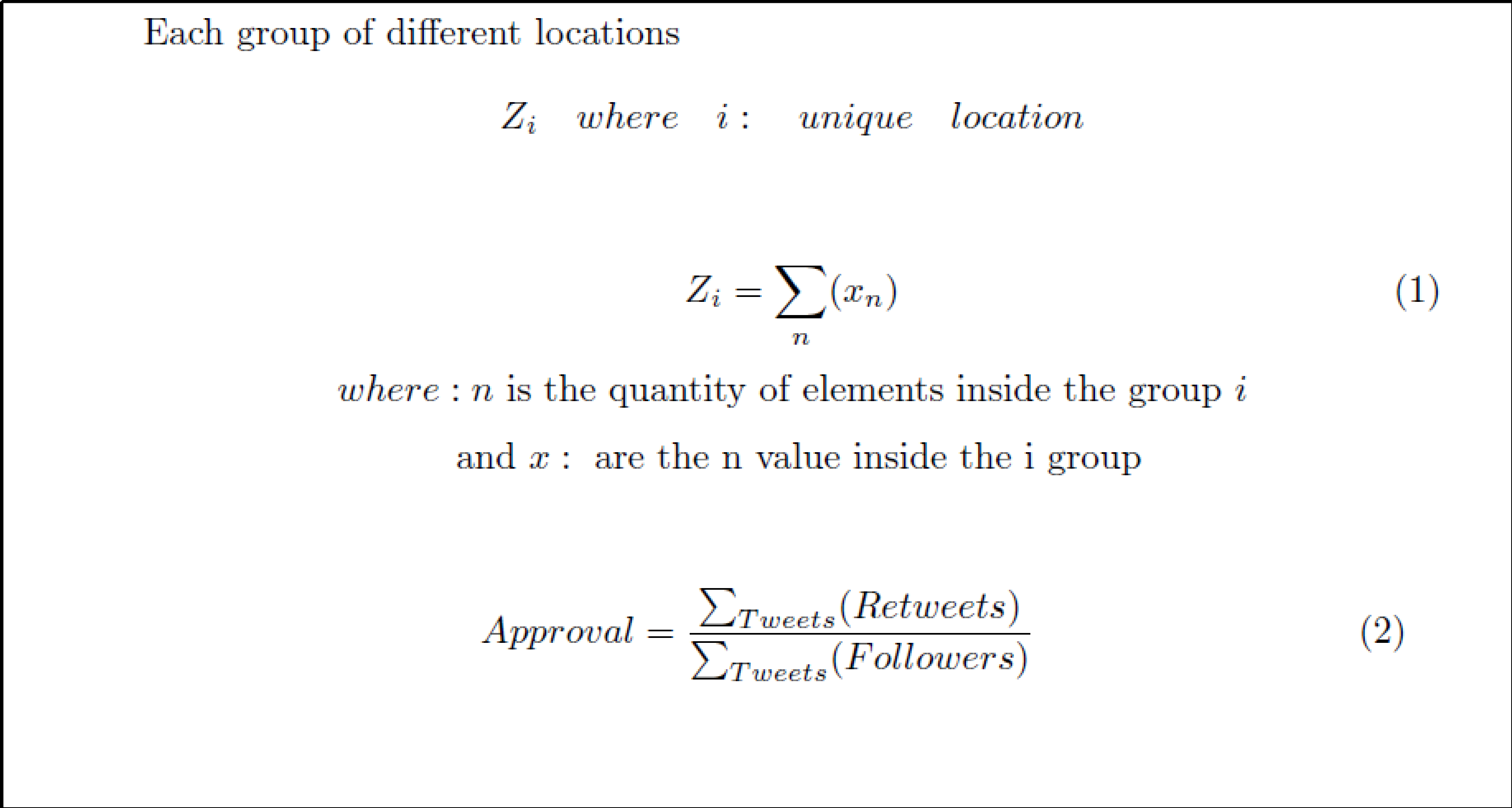


	id	created_at	followers_count	location	favorite_count	retweet_count
1	738276139504963584	2018-09-18 14:13:06	644		0	58
2	2903059609	2018-09-18 14:13:06	302	deep in the heart of texas	0	34302
3	3261609530	2018-09-18 14:13:06	434	Nairobi, Kenya	0	465
4	1034014823594381313	2018-09-18 14:13:05	38		0	0
5	1716299822	2018-09-18 14:13:05	113		0	24592
6	20792010	2018-09-18 14:13:05	11187	Columbus, Ohio	0	0
7	807022147	2018-09-18 14:13:05	1020	Edinburgh, Scotland	0	0

Modeling



- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment



Each group of different locations

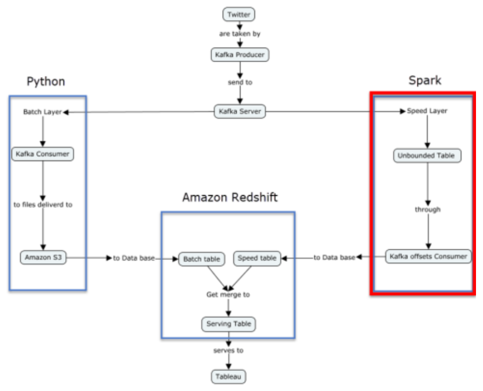
$$Z_i \quad where \quad i: \quad unique \quad location$$

$$Z_i = \sum_n (x_n) \tag{1}$$

$where: n$ is the quantity of elements inside the group i

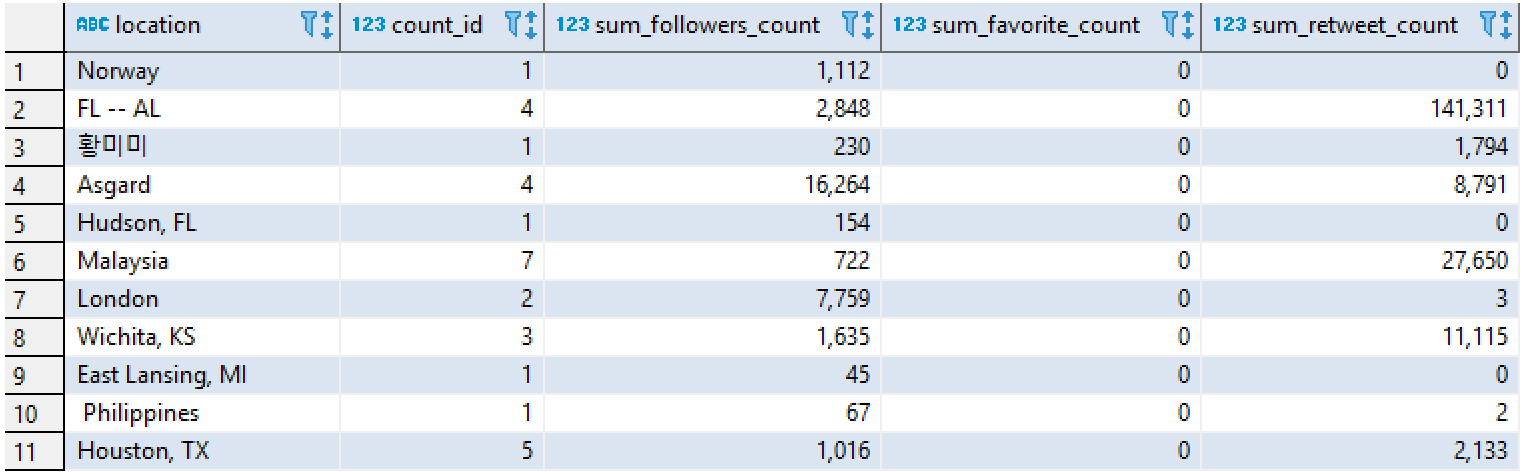
and $x:$ are the n value inside the i group

$$Approval = \frac{\sum_{Tweets}(Retweets)}{\sum_{Tweets}(Followers)} \tag{2}$$



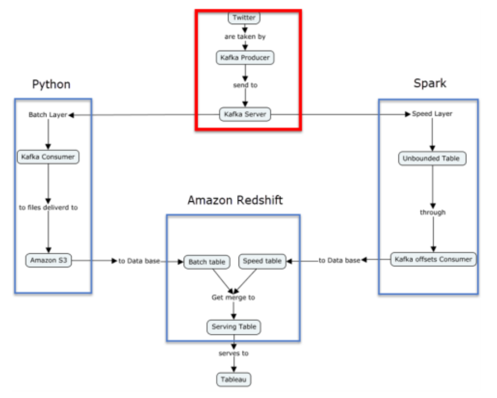
Reagrouped by location (Speed layer Table)

- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

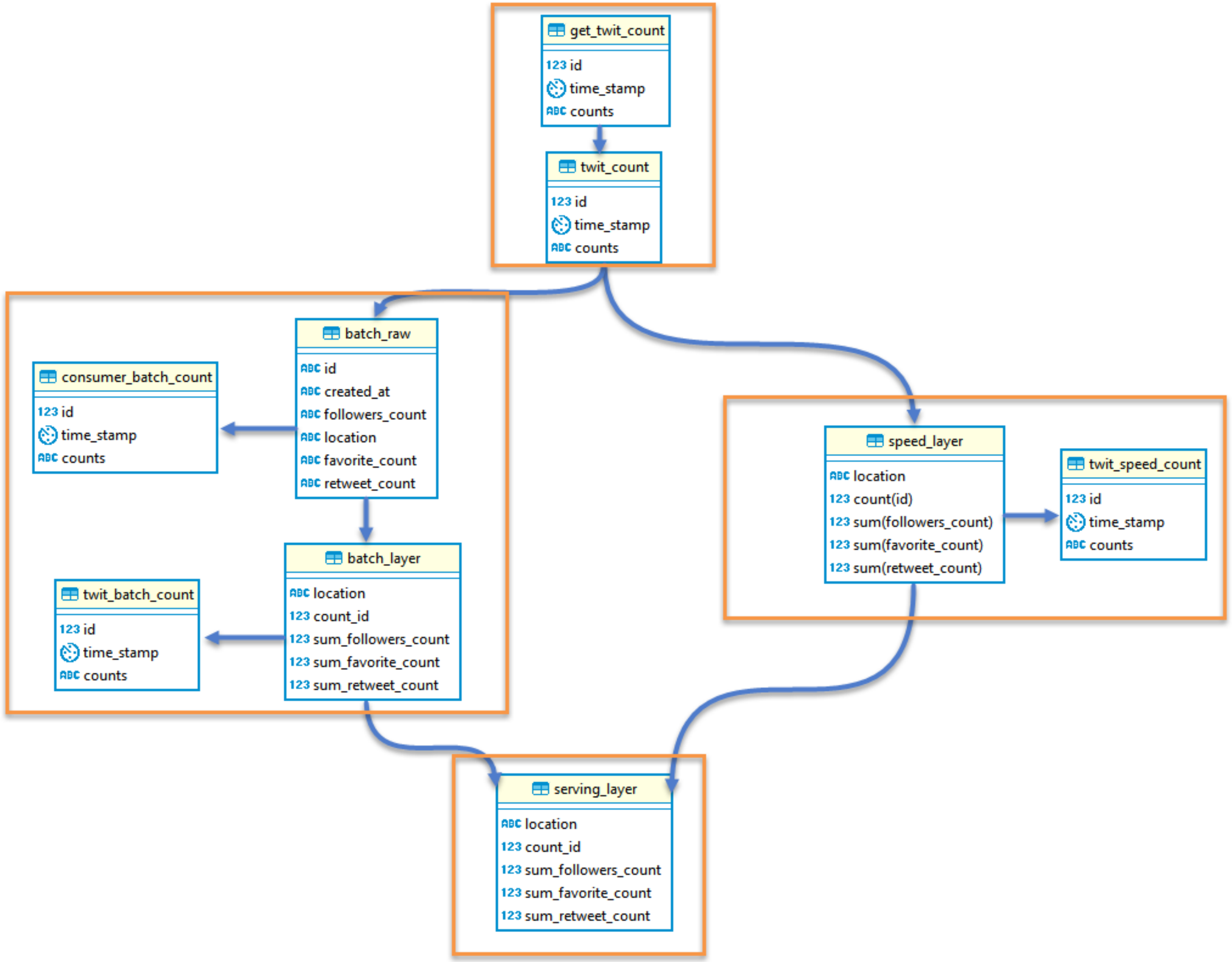


	ABC location	123 count_id	123 sum_followers_count	123 sum_favorite_count	123 sum_retweet_count
1	Norway	1	1,112	0	0
2	FL -- AL	4	2,848	0	141,311
3	촬미미	1	230	0	1,794
4	Asgard	4	16,264	0	8,791
5	Hudson, FL	1	154	0	0
6	Malaysia	7	722	0	27,650
7	London	2	7,759	0	3
8	Wichita, KS	3	1,635	0	11,115
9	East Lansing, MI	1	45	0	0
10	Philippines	1	67	0	2
11	Houston, TX	5	1,016	0	2,133

Data base diagram



- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment


get_twit_count
123 id
time_stamp
ABC counts

twit_count
123 id
time_stamp
ABC counts

consumer_batch_count
123 id
time_stamp
ABC counts

batch_raw
ABC id
ABC created_at
ABC followers_count
ABC location
ABC favorite_count
ABC retweet_count

twit_batch_count
123 id
time_stamp
ABC counts

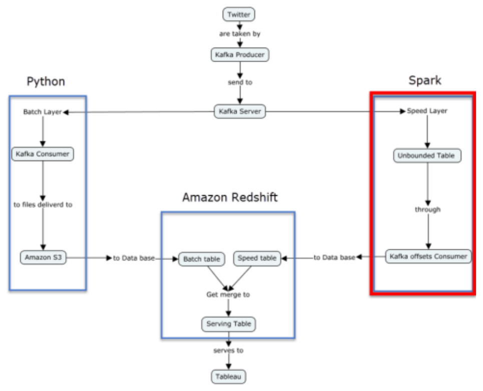
batch_layer
ABC location
123 count_id
123 sum_followers_count
123 sum_favorite_count
123 sum_retweet_count

speed_layer
ABC location
123 count(id)
123 sum(followers_count)
123 sum(favorite_count)
123 sum(retweet_count)

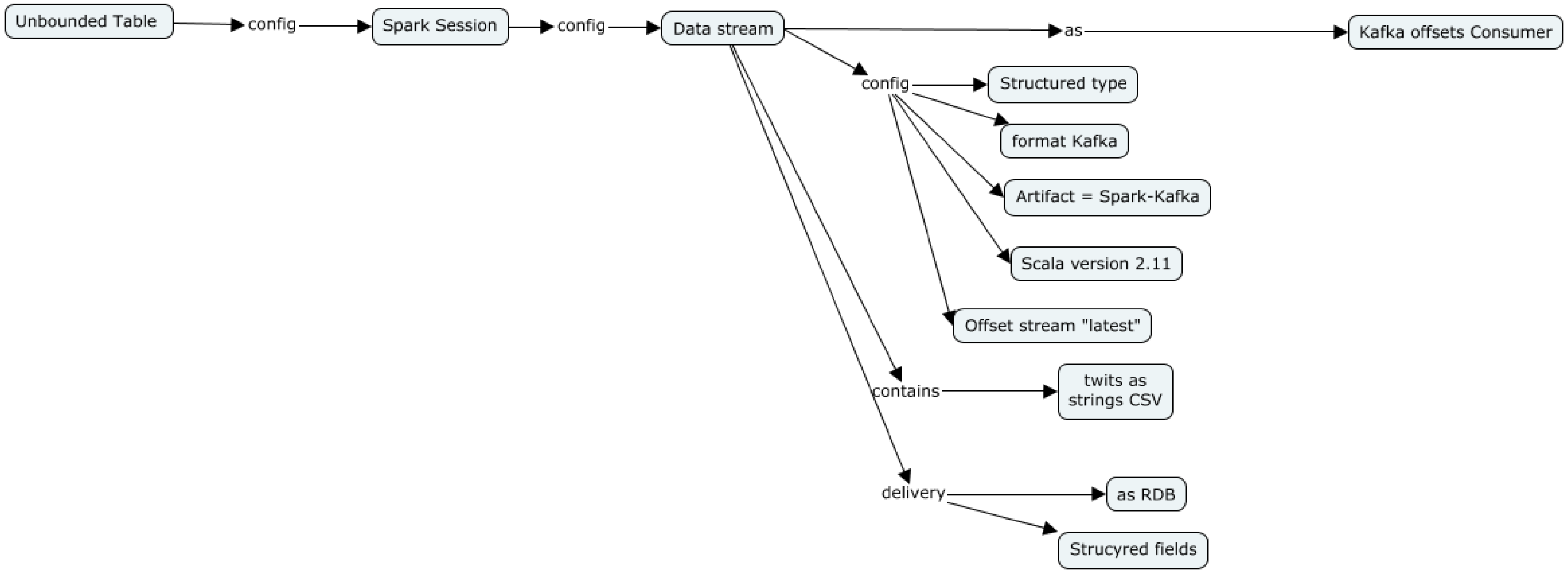
twit_speed_count
123 id
time_stamp
ABC counts

serving_layer
ABC location
123 count_id
123 sum_followers_count
123 sum_favorite_count
123 sum_retweet_count

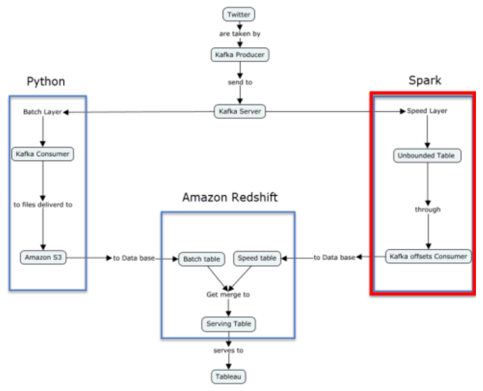
Speed Layer - Spark streaming framework (at-most-once)



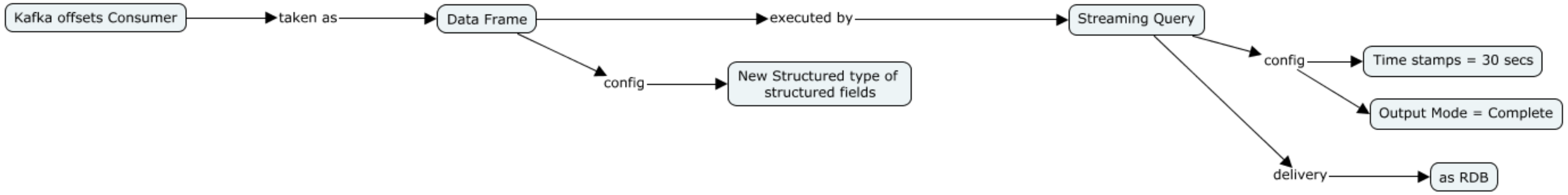
- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment



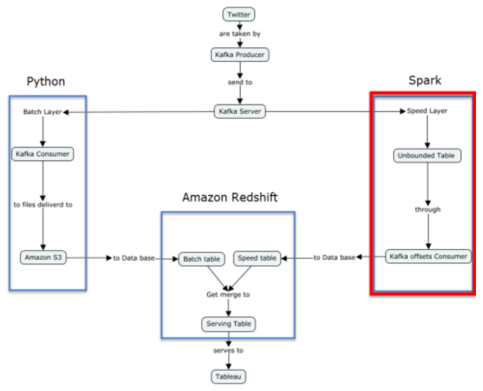
Incremental Query



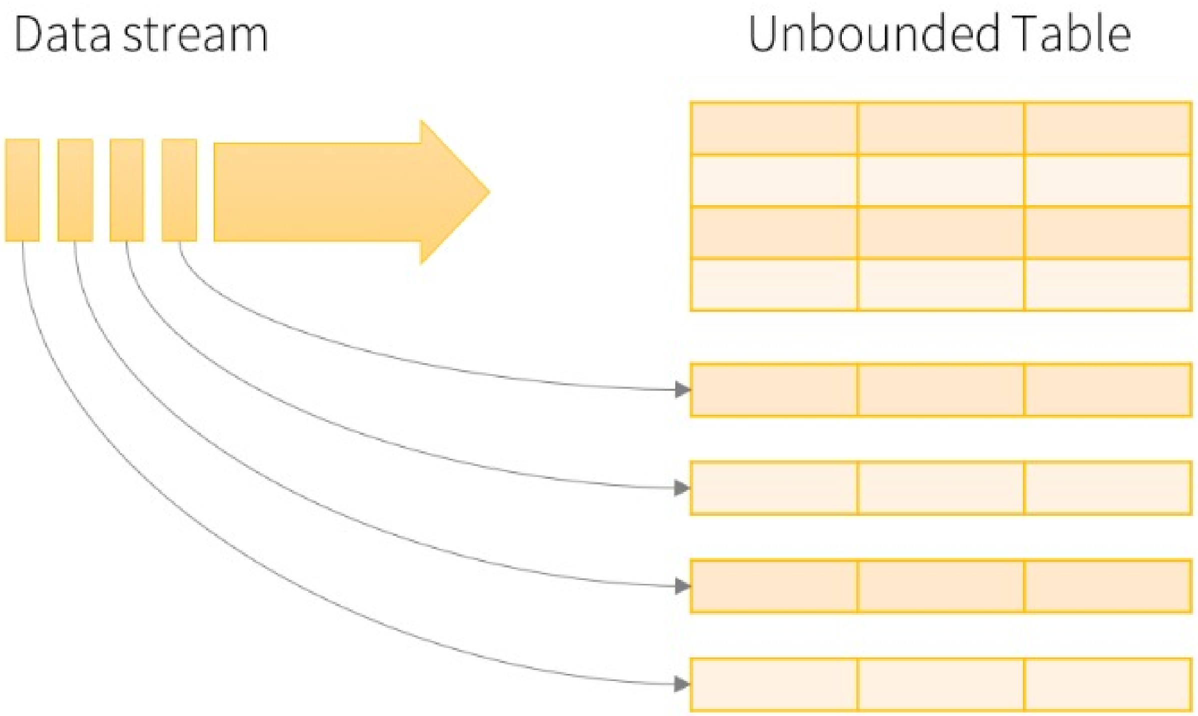
- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment



Incremental Query



- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

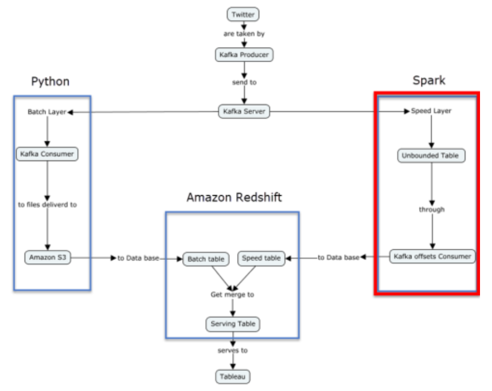


new data in the data stream

=

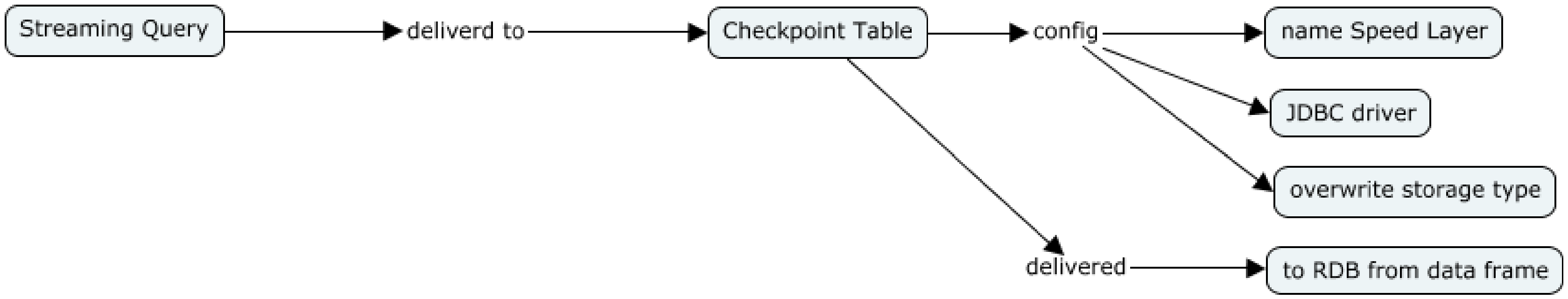
new rows appended to a unbounded table

Data stream as an unbounded table

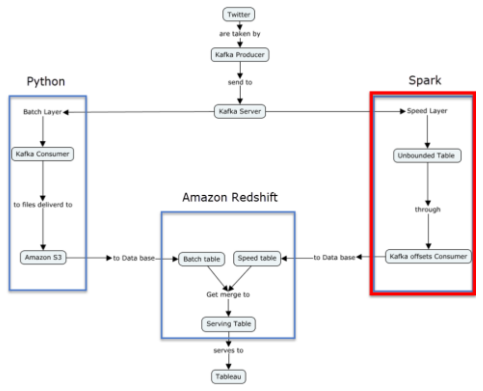


Speed layer – Checkpoint table

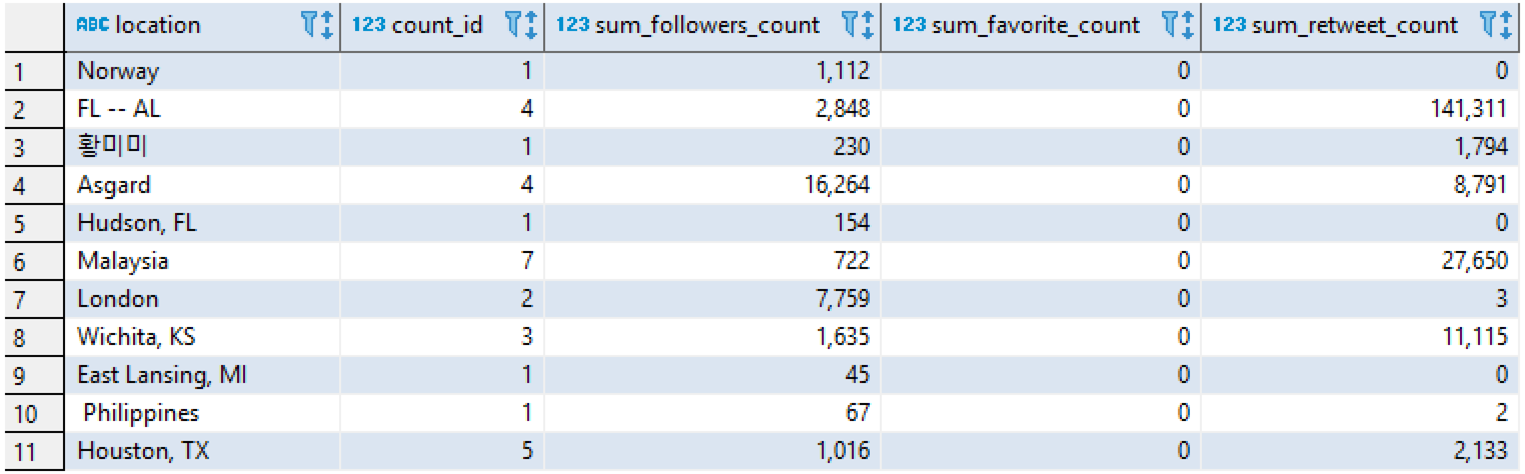
- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment



Speed layer Table

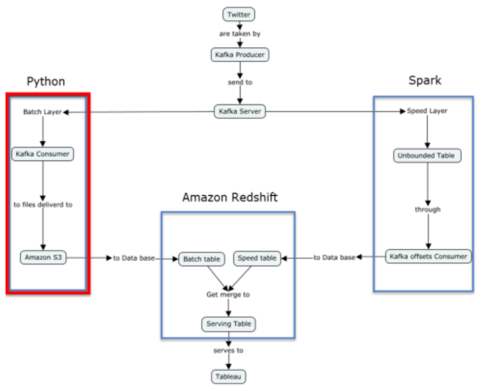


- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

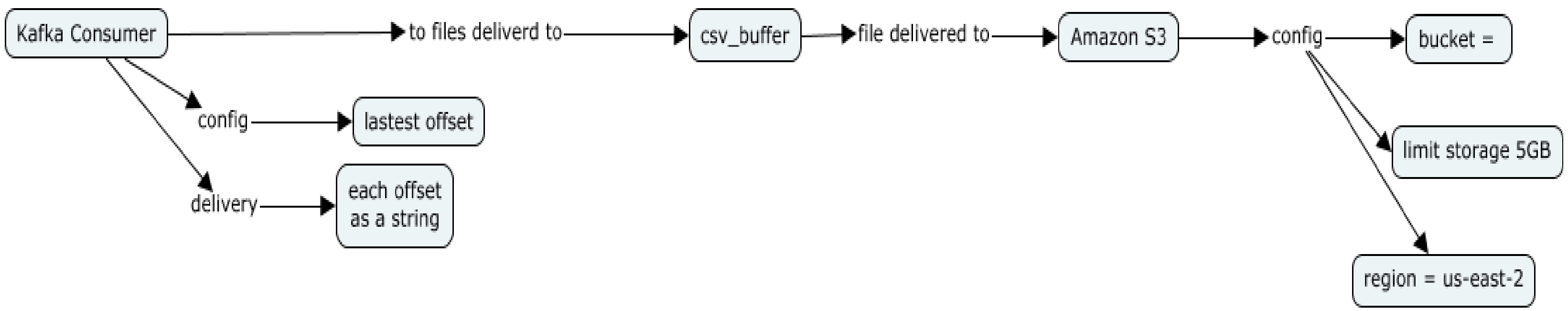


	ABC location	123 count_id	123 sum_followers_count	123 sum_favorite_count	123 sum_retweet_count
1	Norway	1	1,112	0	0
2	FL -- AL	4	2,848	0	141,311
3	활미미	1	230	0	1,794
4	Asgard	4	16,264	0	8,791
5	Hudson, FL	1	154	0	0
6	Malaysia	7	722	0	27,650
7	London	2	7,759	0	3
8	Wichita, KS	3	1,635	0	11,115
9	East Lansing, MI	1	45	0	0
10	Philippines	1	67	0	2
11	Houston, TX	5	1,016	0	2,133

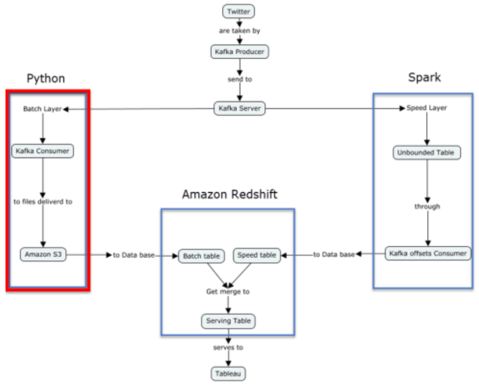
Batch layer – file container S3



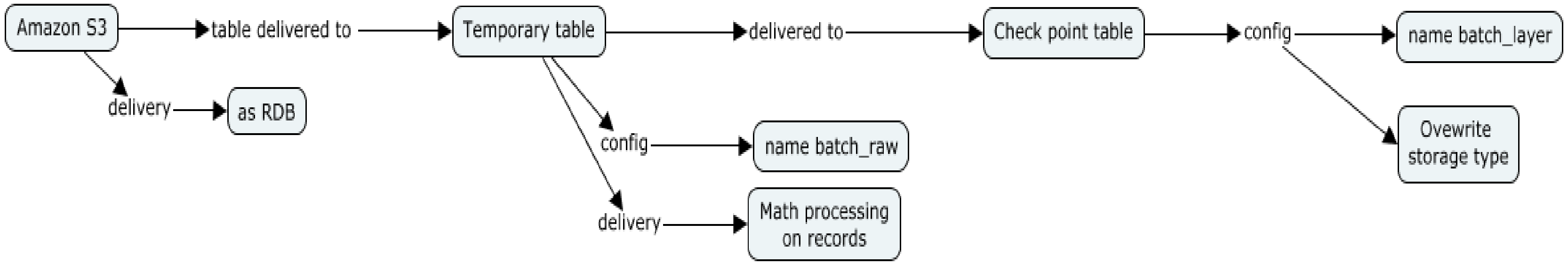
- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment



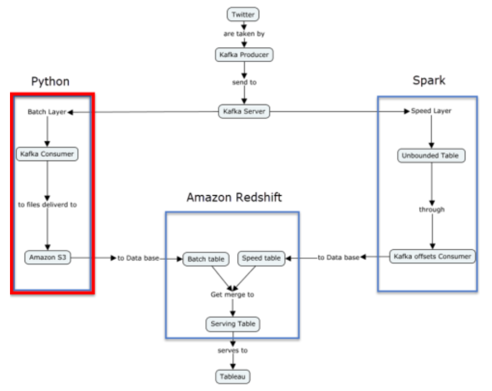
Batch layer – Temporary and chepoint tables



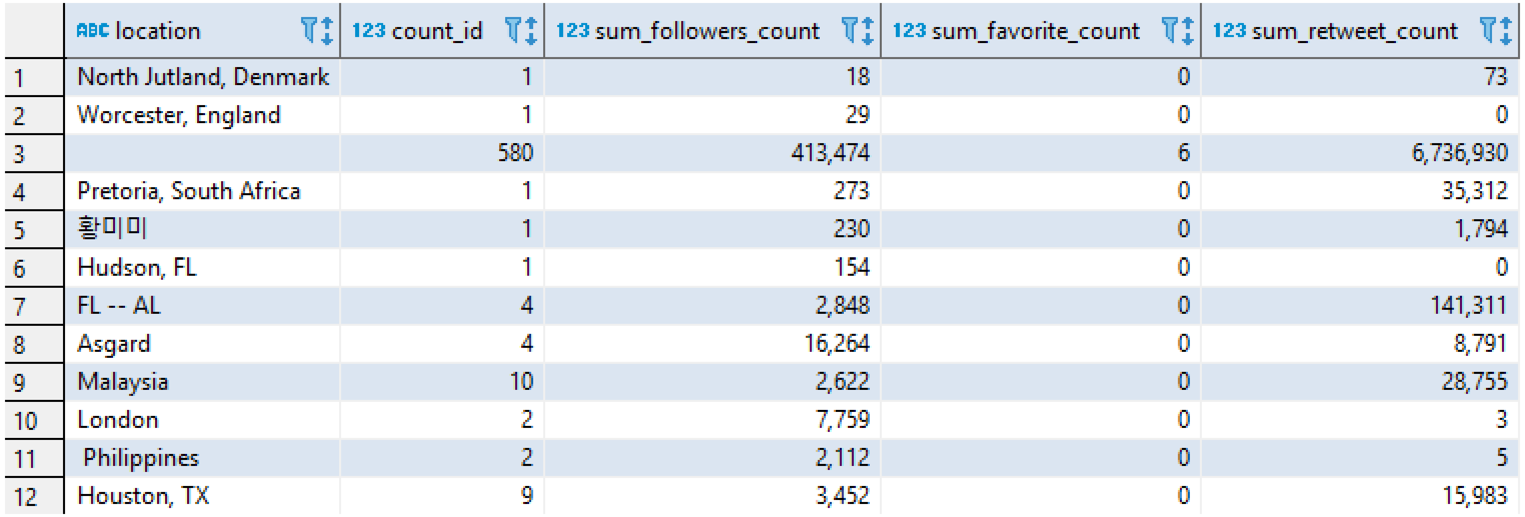
- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment



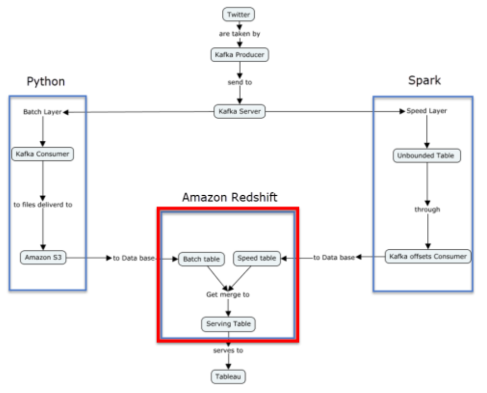
Batch Layer table



- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

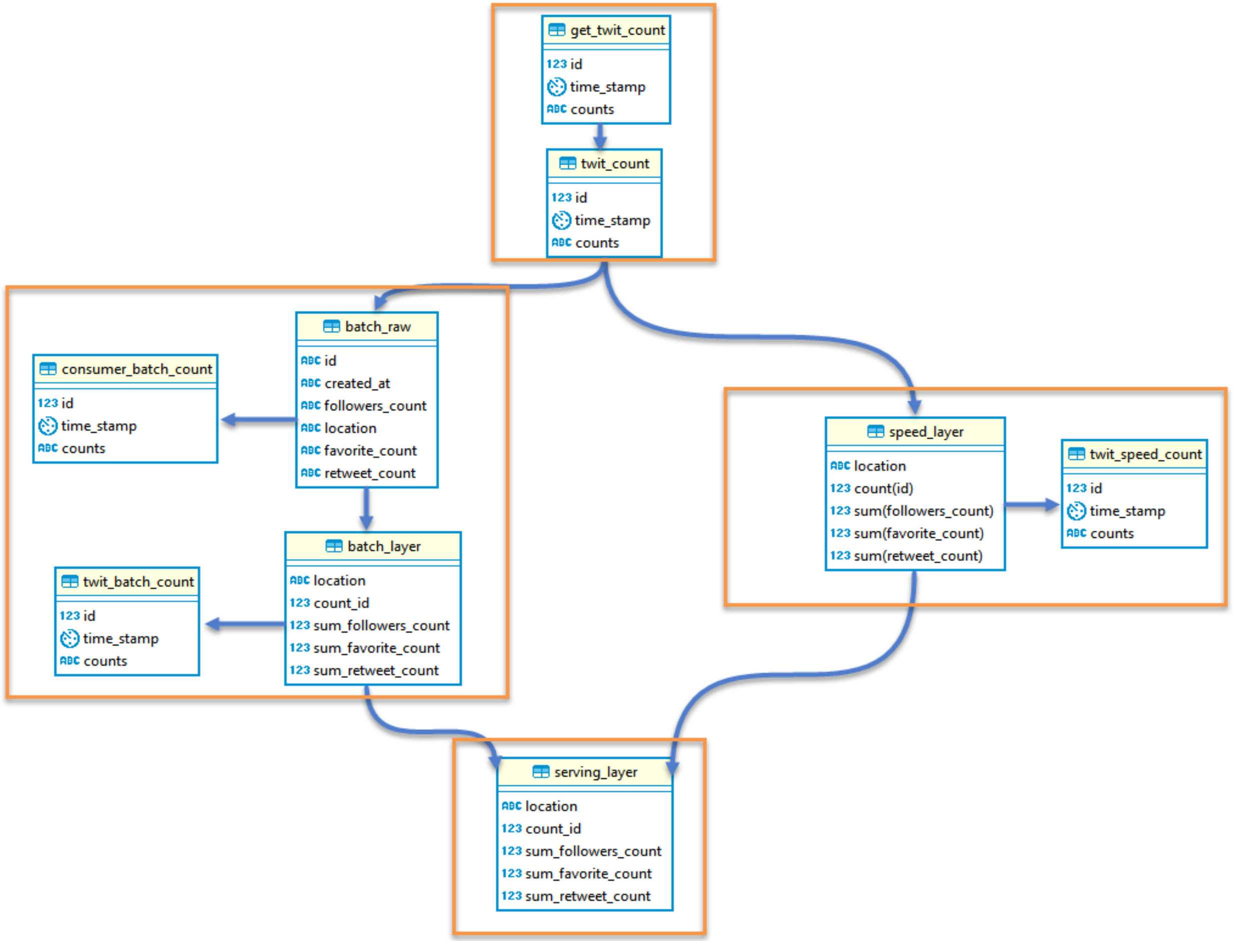


	ABC location	123 count_id	123 sum_followers_count	123 sum_favorite_count	123 sum_retweet_count
1	North Jutland, Denmark	1	18	0	73
2	Worcester, England	1	29	0	0
3		580	413,474	6	6,736,930
4	Pretoria, South Africa	1	273	0	35,312
5	황미미	1	230	0	1,794
6	Hudson, FL	1	154	0	0
7	FL -- AL	4	2,848	0	141,311
8	Asgard	4	16,264	0	8,791
9	Malaysia	10	2,622	0	28,755
10	London	2	7,759	0	3
11	Philippines	2	2,112	0	5
12	Houston, TX	9	3,452	0	15,983



Serving layer table and system counters to valutate

- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment


get_twit_count
123 id
time_stamp
ABC counts

twit_count
123 id
time_stamp
ABC counts

consumer_batch_count
123 id
time_stamp
ABC counts

batch_raw
ABC id
ABC created_at
ABC followers_count
ABC location
ABC favorite_count
ABC retweet_count

twit_batch_count
123 id
time_stamp
ABC counts

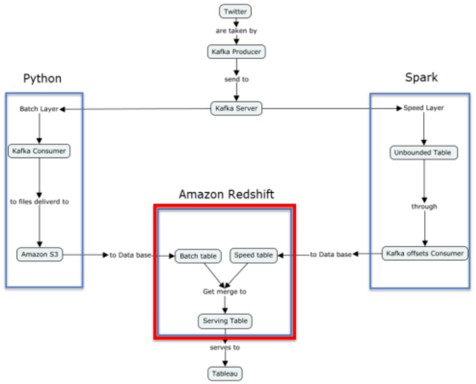
batch_layer
ABC location
123 count_id
123 sum_followers_count
123 sum_favorite_count
123 sum_retweet_count

speed_layer
ABC location
123 count(id)
123 sum(followers_count)
123 sum(favorite_count)
123 sum(retweet_count)

twit_speed_count
123 id
time_stamp
ABC counts

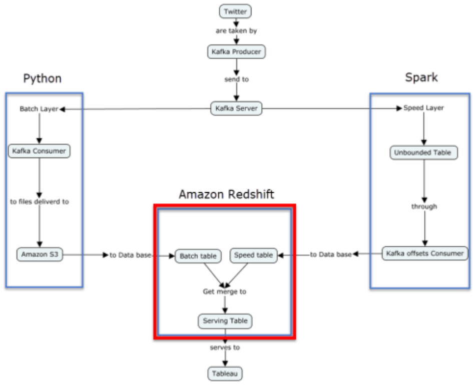
serving_layer
ABC location
123 count_id
123 sum_followers_count
123 sum_favorite_count
123 sum_retweet_count

Tableau «Approval index» Deployment



- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

Redshift Throughput



Twitter

are taken by

Kafka Producer

send to

Kafka Server

Python

Batch Layer

Kafka Consumer

to files deliverd to

Amazon S3

to Data base

Spark

Speed Layer

Unbounded Table

through

Kafka offsets Consumer

to Data base

Amazon Redshift

Batch table

Speed table

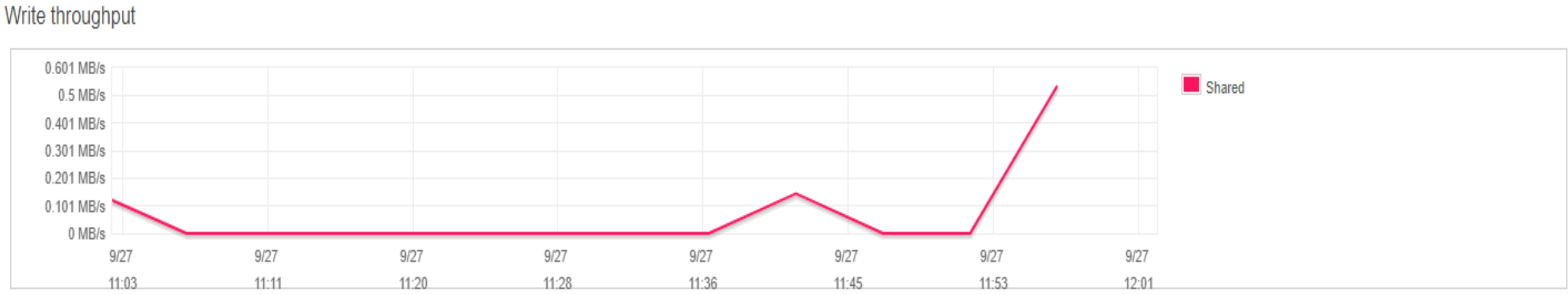
Get merge to

Serving Table

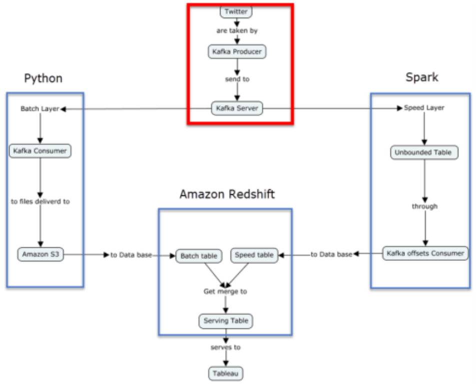
serves to

Tableau

- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

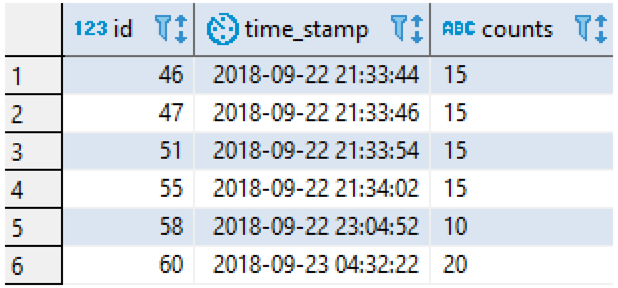


Tweet count table and evaluation parameters



- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

Refreshing of the server period: $\frac{1}{20}\frac{1}{sec}$



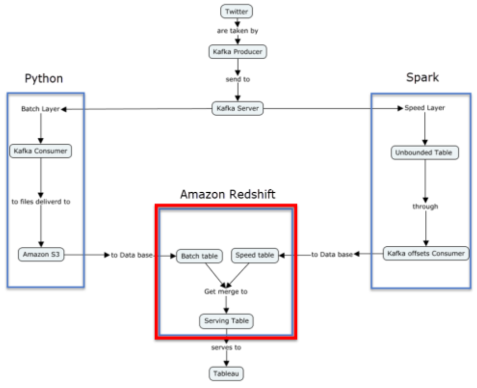
	123 id	time_stamp	ABC counts
1	46	2018-09-22 21:33:44	15
2	47	2018-09-22 21:33:46	15
3	51	2018-09-22 21:33:54	15
4	55	2018-09-22 21:34:02	15
5	58	2018-09-22 23:04:52	10
6	60	2018-09-23 04:32:22	20

Maximum Quantity of tweets in a request: $100 \quad tweets$

	Tweet Count
***1 Minute (morning)**	70
***1 Minute (night)**	284

*average of 5 throughput values

Batch table and Speed table merged into Serving Table



- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

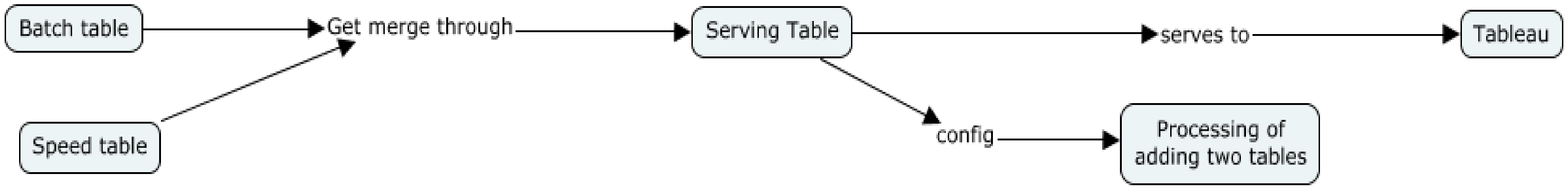
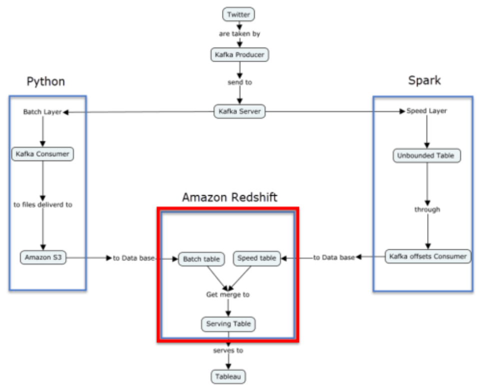


Tableau «Approval index» Deployment



- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

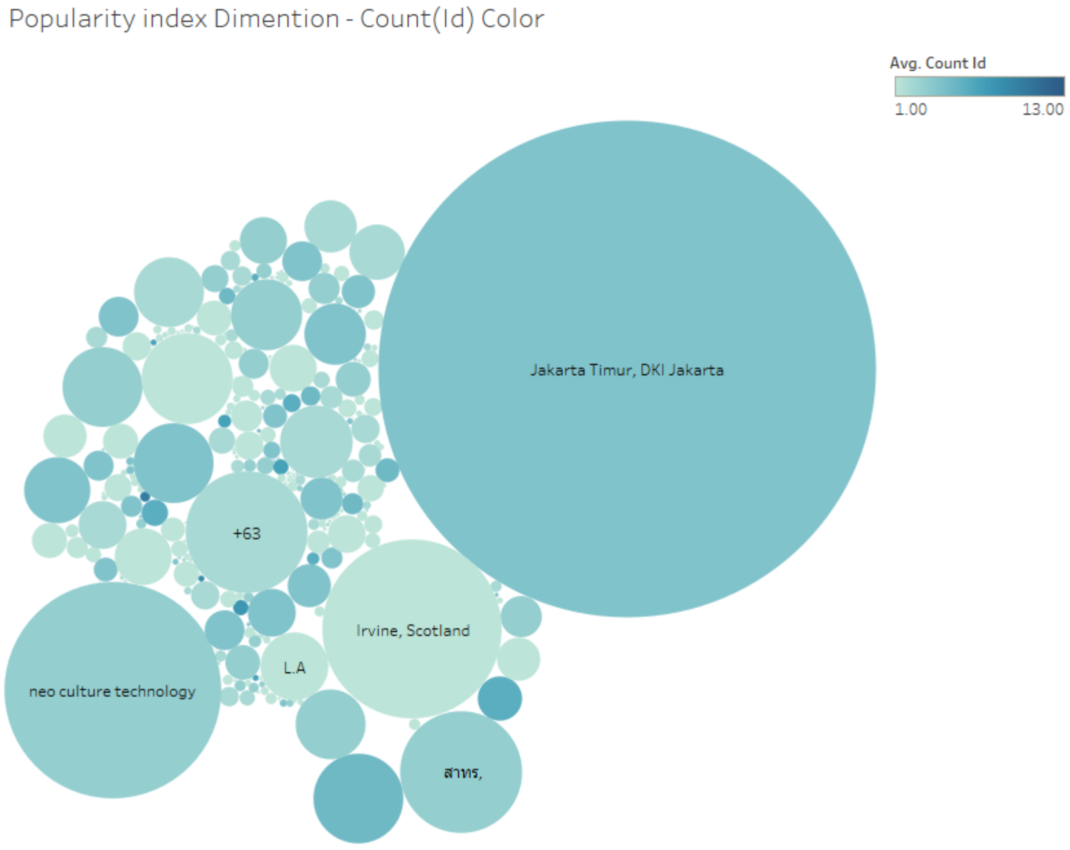
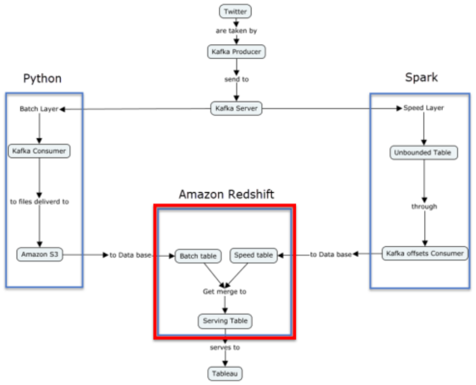

Popularity index Dimention - Count(Id) Color

Tableau «Approval index» Deployment



- Business Understanding
- Data Understanding: Data Discovery
- Data understanding: Data Acquisition and Data Exploration
- Data Preparation and Modelling
- Evaluation
- Presentation and Deployment

- The implemented system represents a kind of visual stethoscope, thanks to real time visualization, is able to give the public opinion sensibility about a certain topic.
- Is more desirable the search() method than the streaming filtered, due to the capability to search in different languages, if the business hasn't their own data semantics system.
- The lambda architecture pipeline is being wasted, due to the implementation of a system with a throughput of 1KB per minute and with a data acquisition of historical data.
- The Business goal makes that the desired Kafka streaming processing has to be "at-most-once" implementation, making no replication, but with a probability of losing data.
- **Further steps** could be the use of **data wrangling** systems like **Google Fusion Tables, OpenRefine** and others, to only keep track of meaningful locations; and the implementation of Twitter streaming filter with a system of data semantics, if the desired word, to look for, exceeds the Rate limit, of request per minute, that the search() method offer.