BigFrame

A Benchmark Tool for Big Data Analytics

Andy
Hong Kong PolyU

What is BigFrame

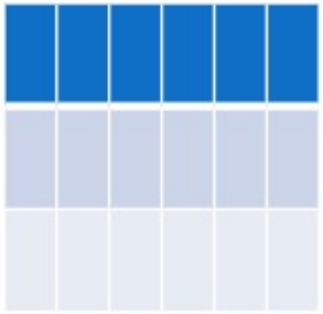
- A benchmark tool that can:
  - generate a large amount of data (PB of data)
  - generate a mixture of social network (graph), tweet (text) and e-commerce (relational) data.
  - provide benchmark workflows (hadoop implementation)
- Open source, free to use

How BigFrame helps hadoop

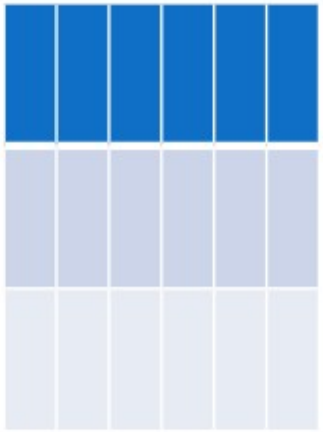
- Testing hadoop:
  - Can it handle multiple data at the same time?
  - What is the bottleneck?
  - How should we tune hadoop's system parameters?

Data Generated

Item

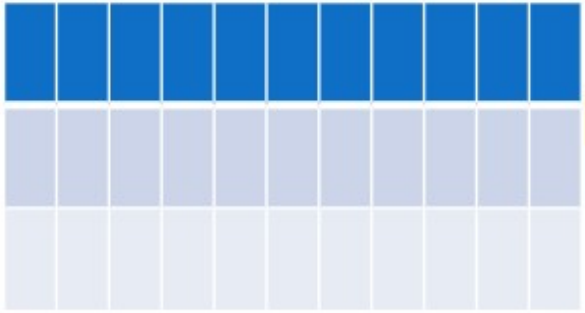


Web_sales

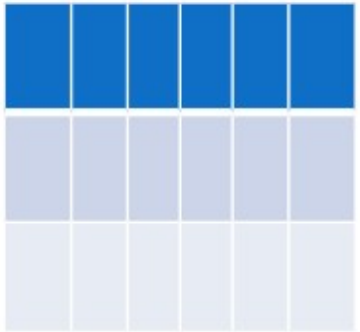


Tweets

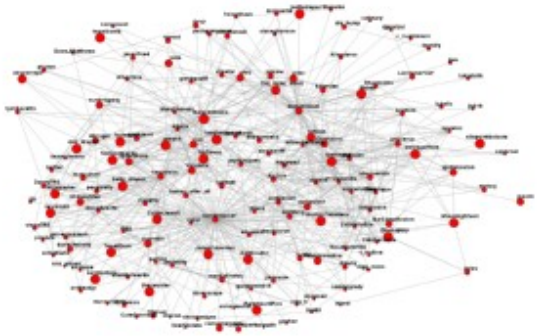
Customer



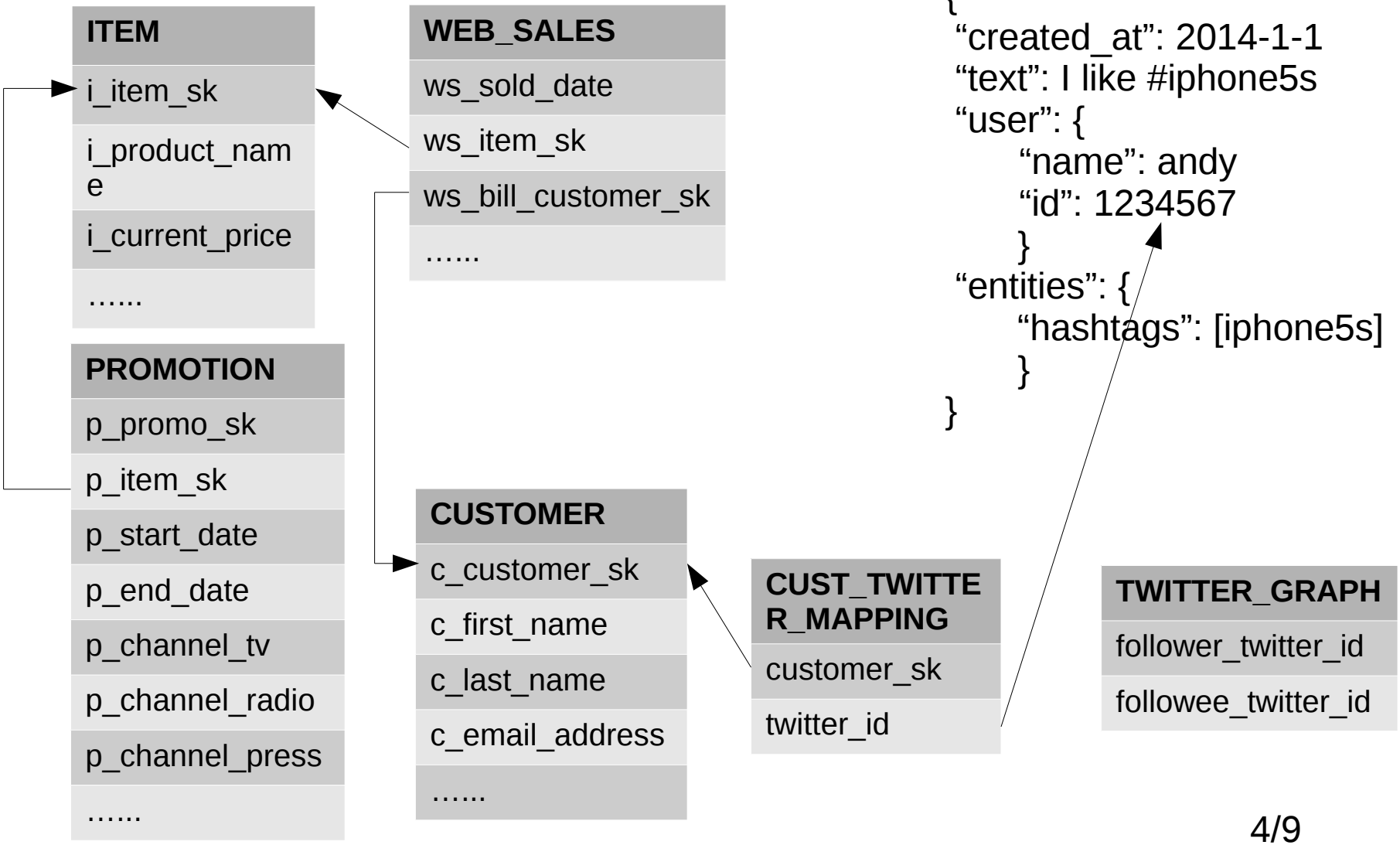
Promotion



Relationships



Data Schema

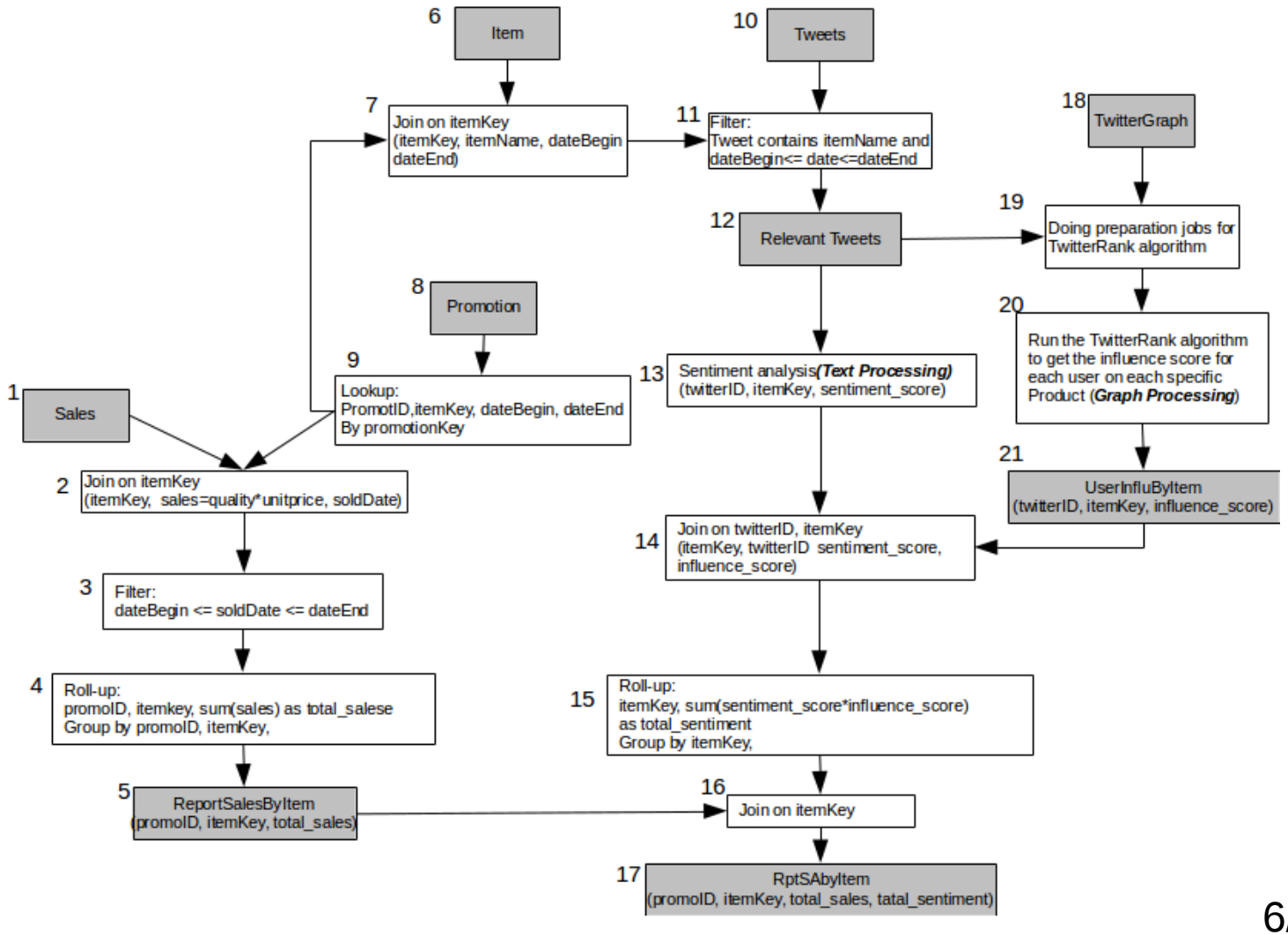

ITEM
i_item_sk
i_product_name
i_current_price
……
WEB_SALES
ws_sold_date
ws_item_sk
ws_bill_customer_sk
……
PROMOTION
p_promo_sk
p_item_sk
p_start_date
p_end_date
p_channel_tv
p_channel_radio
p_channel_press
……
CUSTOMER
c_customer_sk
c_first_name
c_last_name
c_email_address
……
CUST_TWITTER_MAPPING
customer_sk
twitter_id
TWITTER_GRAPH
follower_twitter_id
followee_twitter_id
TWEET:
{
“created_at”: 2014-1-1
“text”: I like #iphone5s
“user”: {
“name”: andy
“id”: 1234567
}
“entities”: {
“hashtags”: [iphone5s]
}
}

Benchmark Workflow

- The benchmark workflow relates the sales report to the sentiment analysis of all promoted products after a set of promotions is ended
  - The example result is:

promo_sk	item_sk	total_sales	sentiment_score
1	10	10,000	45
2	30	100	-67

Benchmark Workflow



Install BigFrame

- Download it from GitHub:
  - git clone https://github.com/bigframeteam/BigFrame
- Run the following command in bigframe's root directory to compile it:
  - *sbt/sbt assembly*

Generate the benchmark data

- Specify the data size by setting the value of property “***bigframe***.***datavolume***” in the “***conf/bigframe-core.xml***” file
- Run the command in the root directory of bigframe:
  - *./bin/datagen -mode datagen*

Run the benchmark workflow

- The implementation is provided: workflow
- Run the command in the root directory of bigframe to start the benchmark:
  - *./bin/qgen -mode runqueries*