Data Engineering for Data Science

Arup Nanda

VP, Data Services

Priceline





booking.com priceline.com kayak.com agoda.com rentalcars.com opentable.com



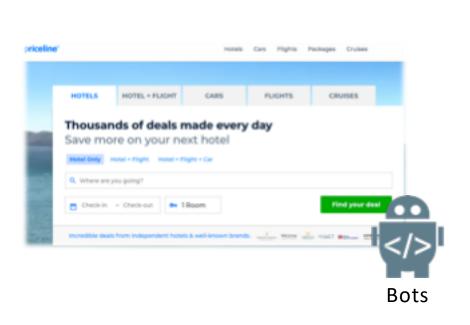
Data Science and Machine Learning

Customer Segmentation Prediction of Behavior

Data Science and Machine Learning

Machine Behavior
Predictive Setting of System
Properties



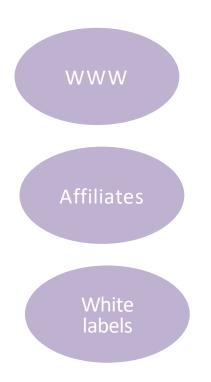


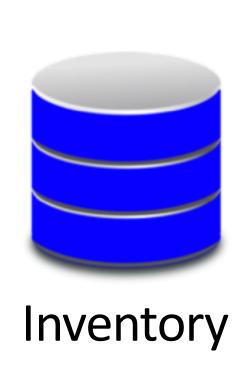


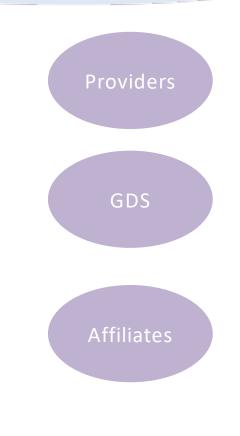
Demand

Technology

Supply







250K Hotels X 200 rooms X 365 days X

Demand

Technology

Supply







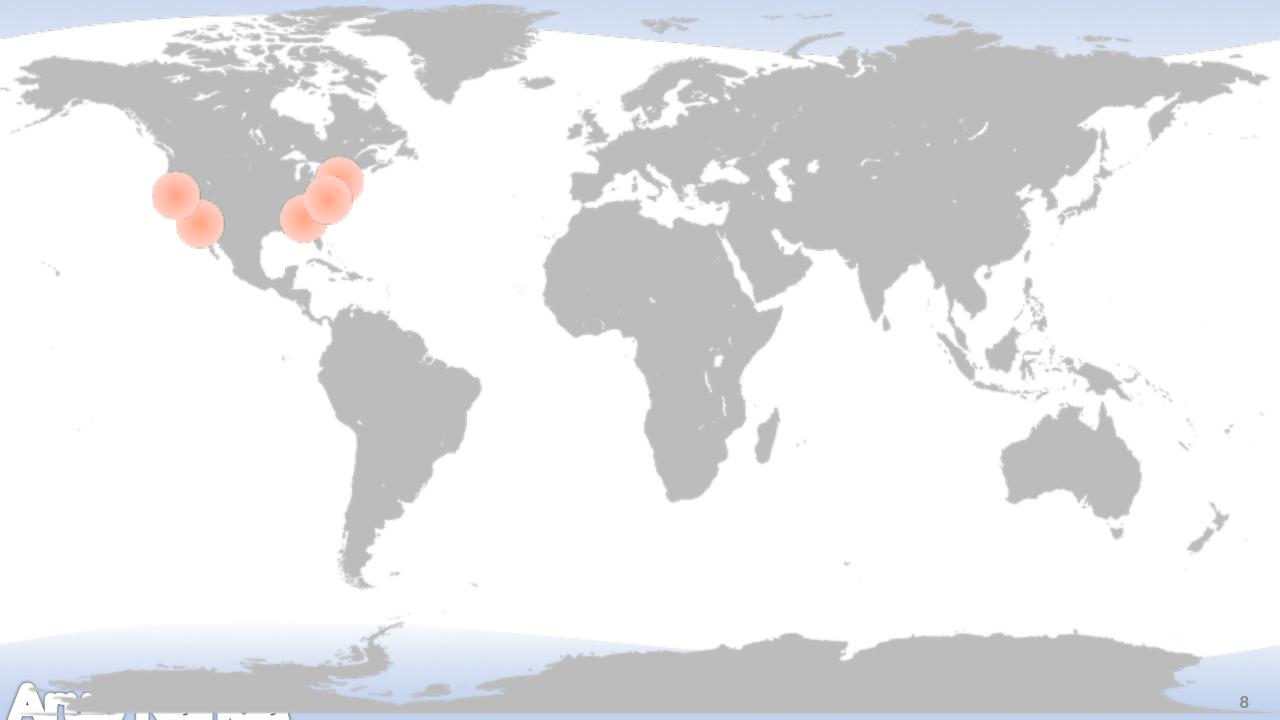


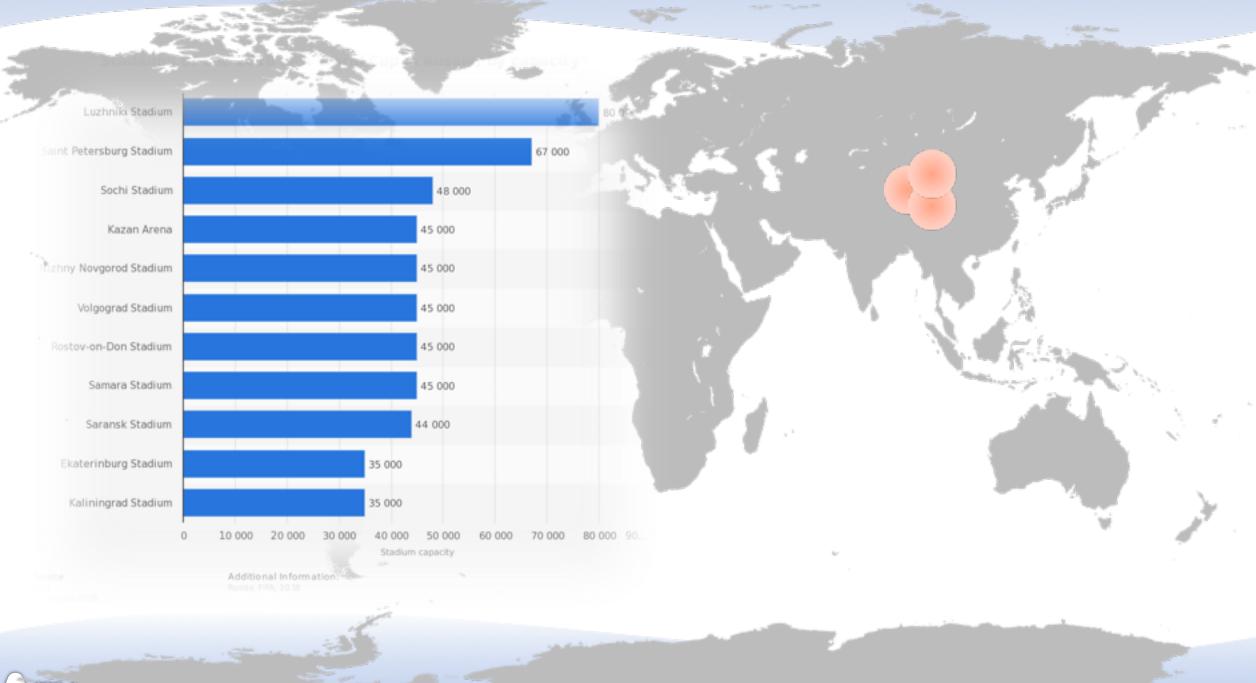
Demand

Technology

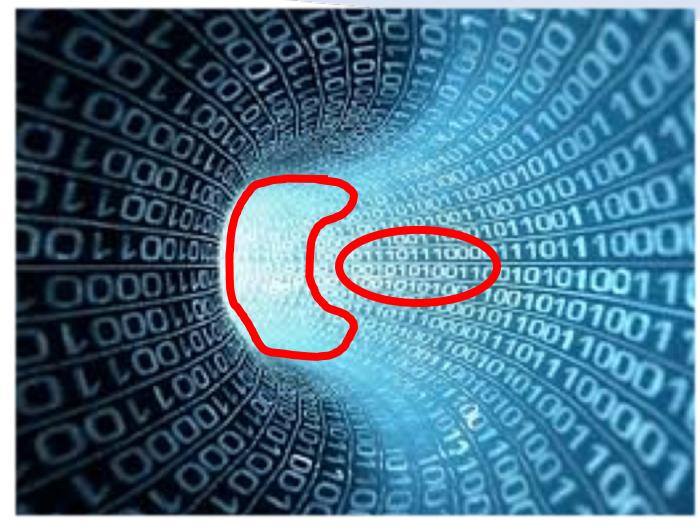
Supply



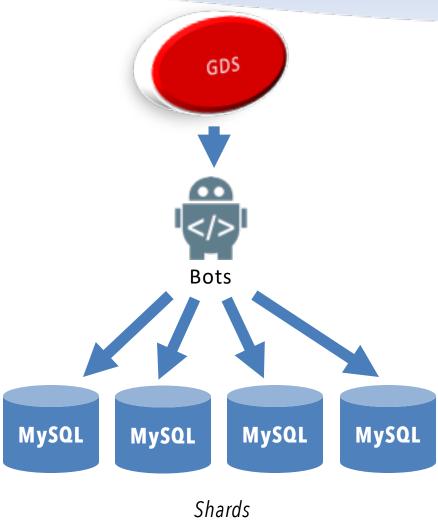


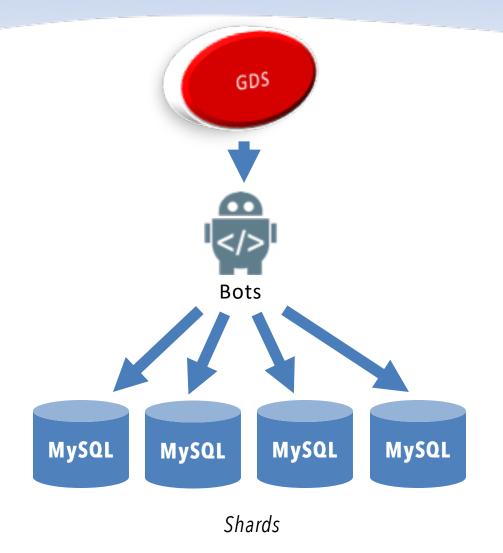




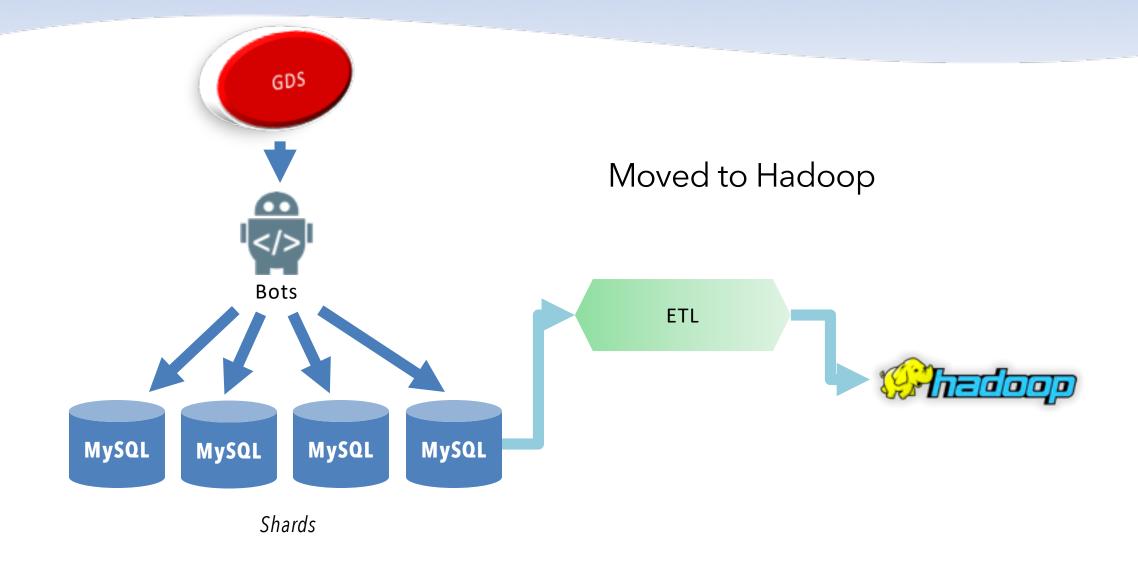


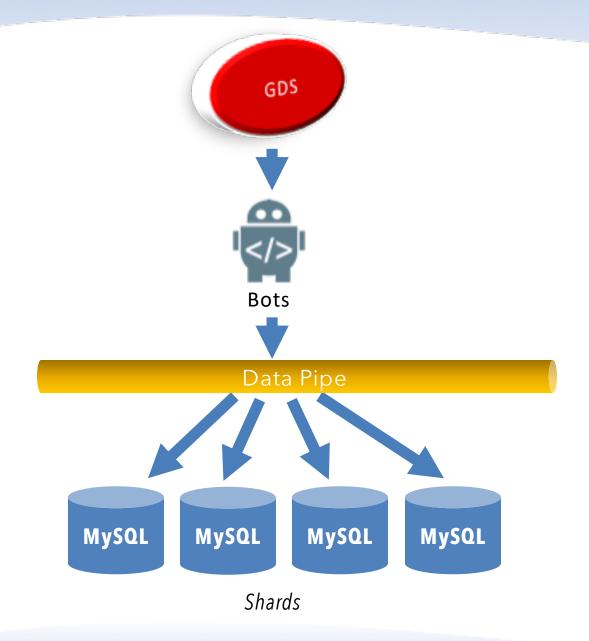
Smart Cache Management



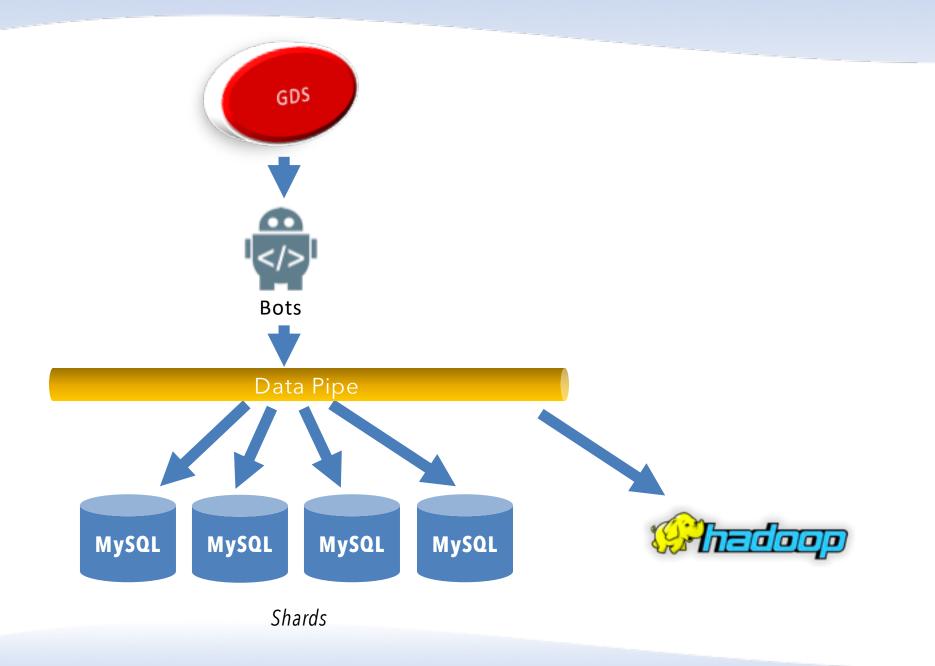


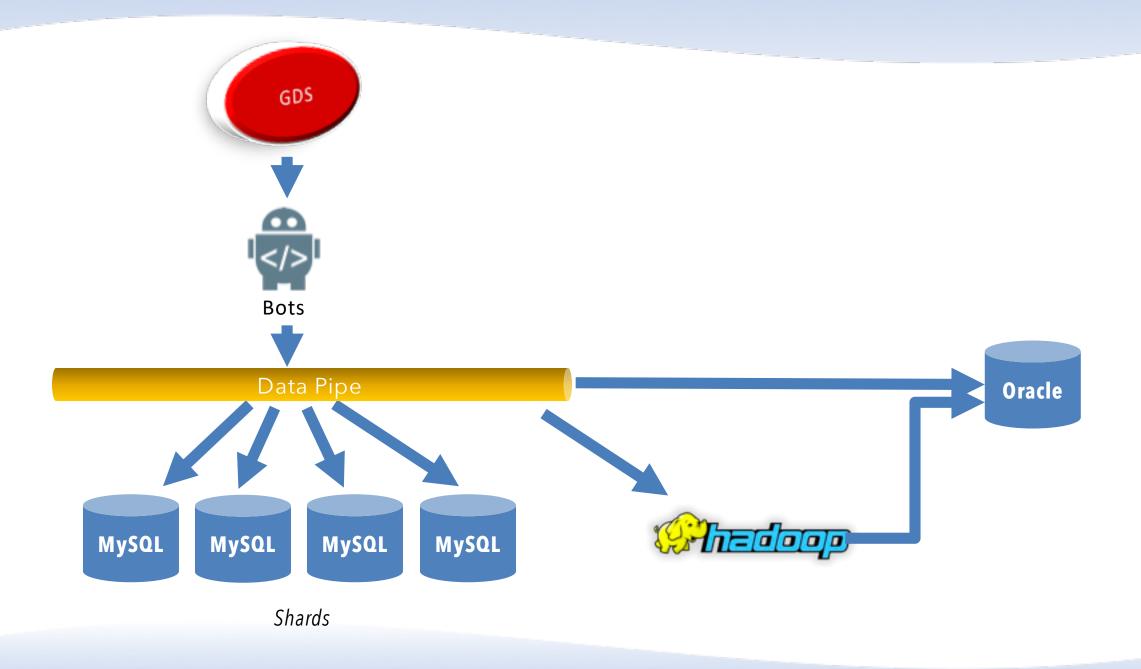
Inventory and Rates stored as **blob**

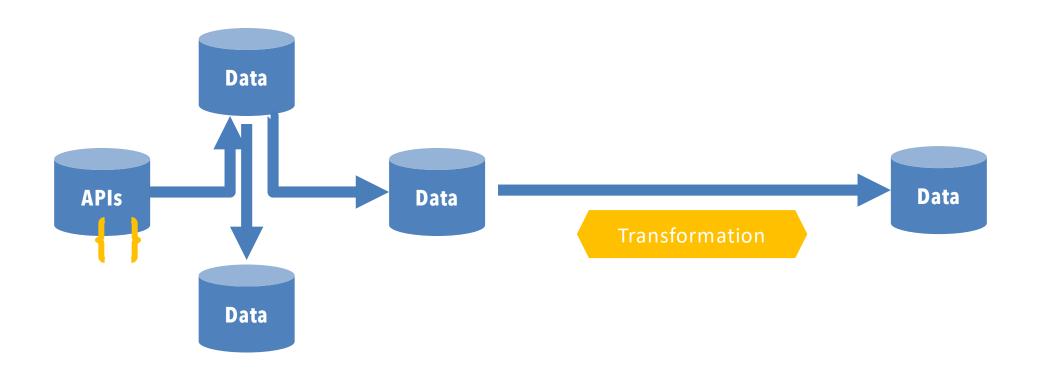


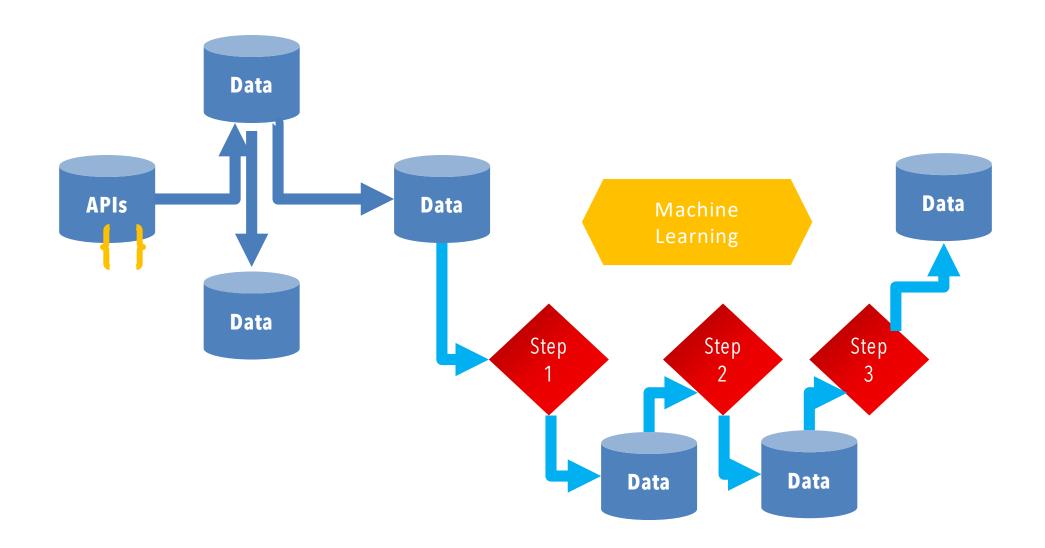


based on Kafka named *Mississipi* has APIs and common libraries







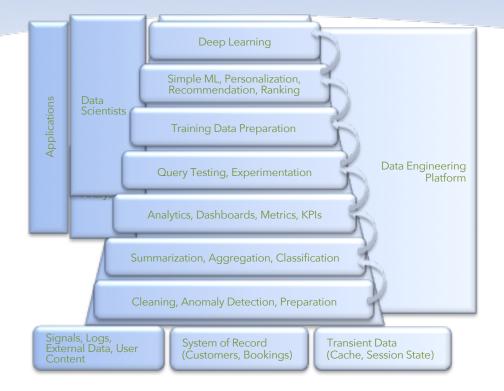




Signals, Logs, External Data, User Content

System of Record (Customers Bookings)

Transient Data (Cache, Session State)



Data Engineering Wishlist

- 1. Repeatable Code
- 2. Easy to transform
- 3. Quick to market
- 4. Scheduling built in
- 5. Customer can monitor
- 6. Analyze the past metrics

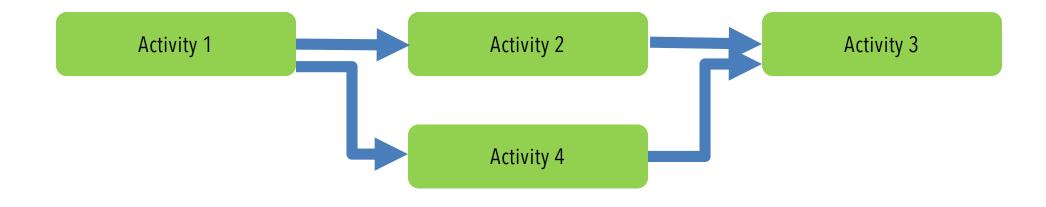
FREE



Apache Airflow



Directed Acyclic Graphs (DAG)



Directed = In one direction only. Can't go back

Acyclic = Must finish. Must be Idempotic





Task Tries

Landing Times

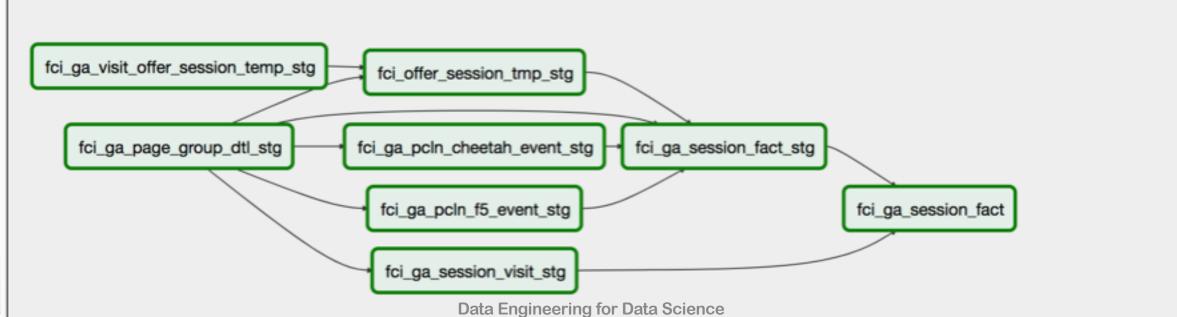
Gantt

Details

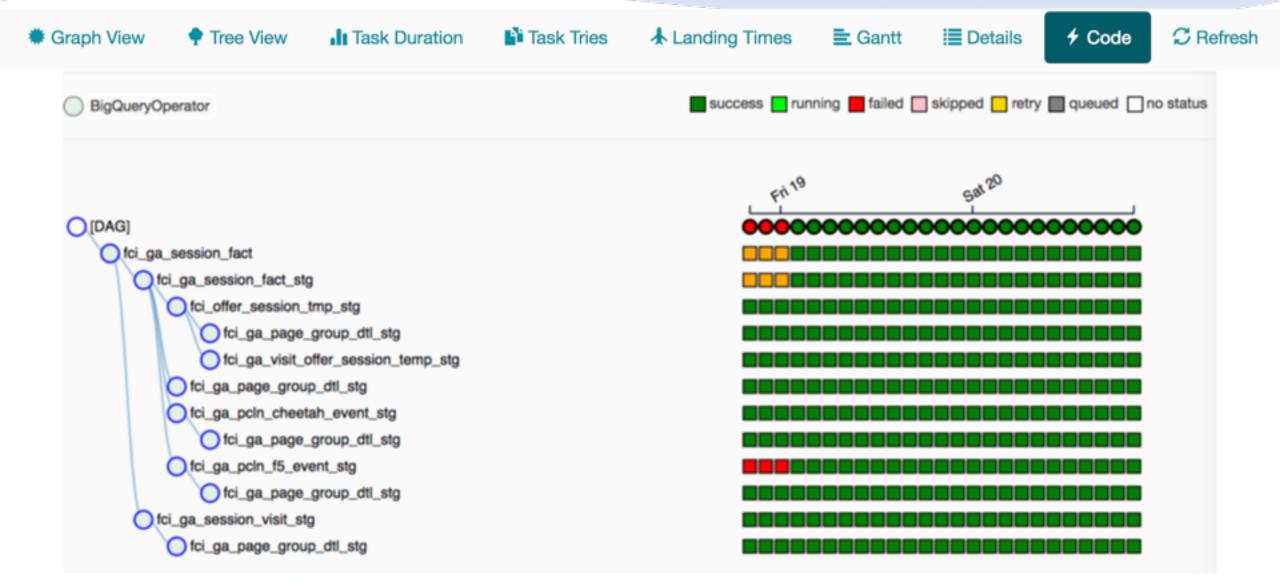
success Run: scheduled_2018-10-20T20:10:00 \$ Layout: Left->Right \$ Go

BigQueryOperator

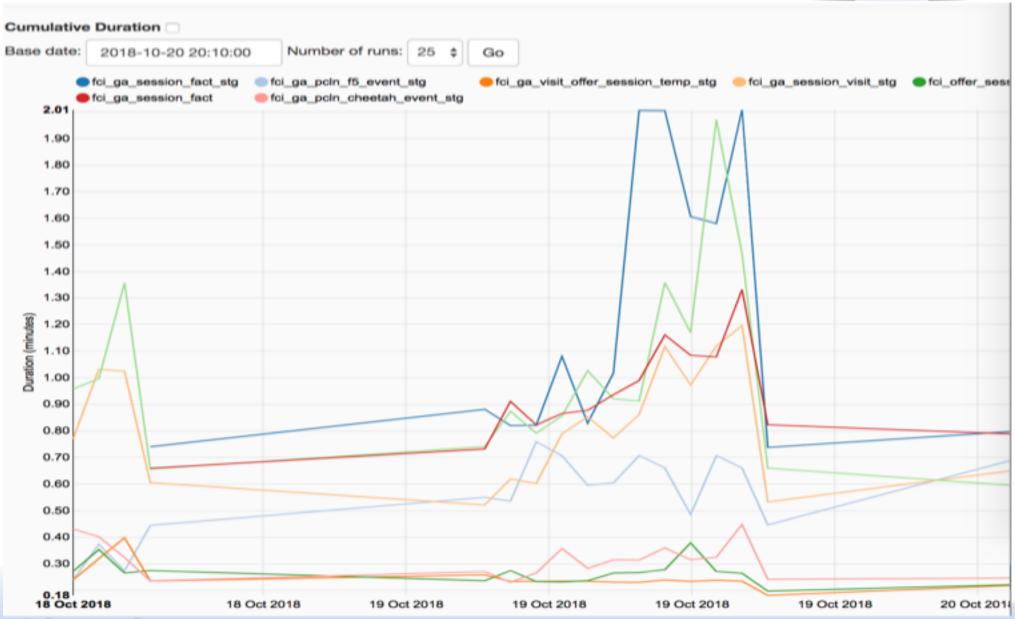
Airflow















Tree View

Task Duration



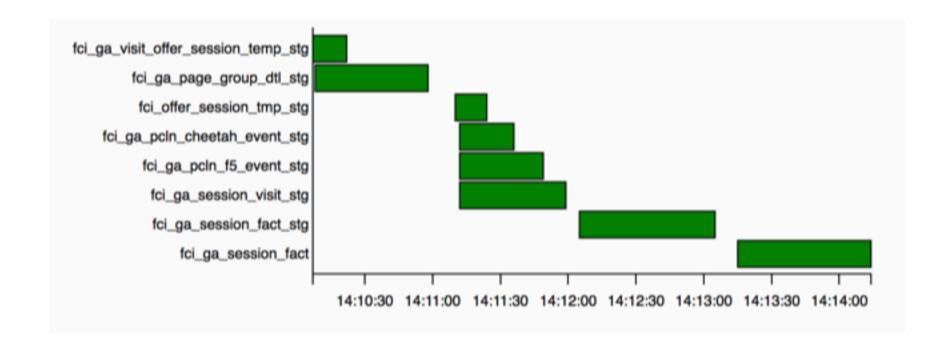
★ Landing Times





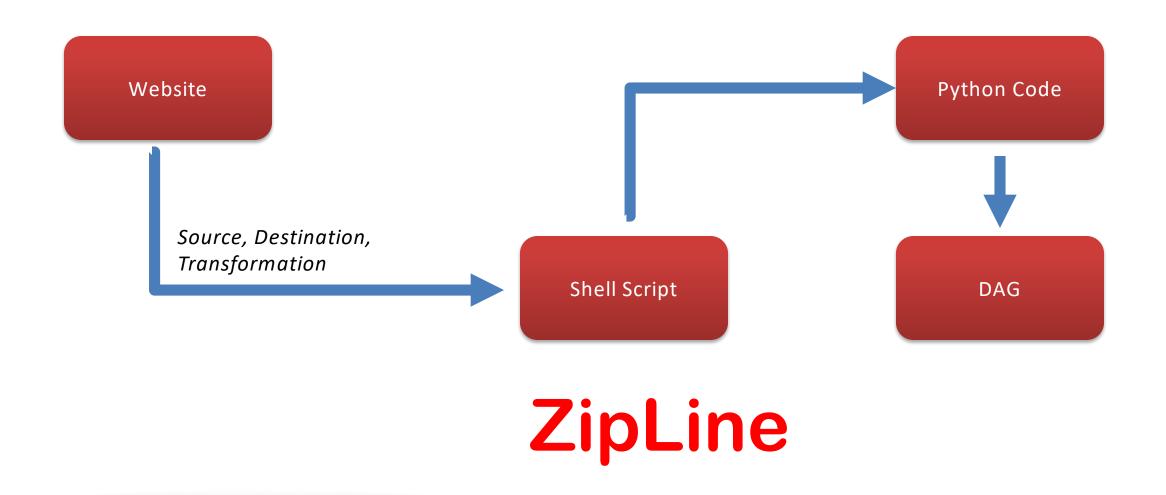








```
import airflow
   from airflow.contrib.operators.bigquery_operator import BigQueryOperator
    from datetime import datetime, timedelta
    import pytz
    def slack_msg(context):
     webhook_url='https://hooks.slack.com/services/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'
     slack_data = {'text': "FCI hourly load did not complete successfully."}
     response = requests.post(
        webhook_url, data=json.dumps(slack_data),
10
11
        headers={'Content-Type': 'application/json'}
12
     return operator.execute(context=context)
13
14
    default_dag_args = {
16
        'start_date': datetime(2018, 9, 27),
17
        'email': 'geetha.rajgopalan@priceline.com',
18
        'email_on_failure': True,
19
        'email_on_retry': False,
20
        'retries': 0,
21
        'retry_delay': timedelta(minutes=10),
22
        'catchup': False
23
        #'on_failure_callback': slack_msg
24 }
25
26
    with airflow.DAG(
27
            'fci_wkflow_hourly_v3',
28
            schedule_interval="10 12-23 * * *",
29
            #schedule_interval=None,
30
            default_args=default_dag_args) as dag:
31
32
      #processDate = datetime.strftime(datetime.now() - timedelta(1), '%Y-%m-%d')
33
      tz = pytz.timezone('US/Eastern')
34
      processDate = datetime.strftime(datetime.now(tz), '%Y-%m-%d')
35
      processDate_nodash = datetime.strftime(datetime.now(tz), '%Y%m%d')
      processTimestamp = datetime.strftime(datetime.now(tz), '%Y-%m-%d %H:%M:%S')
36
37
      fci_ga_page_group_dtl_stg = BigQueryOperator(
38
                                                      Data Engineering for Data Science
```







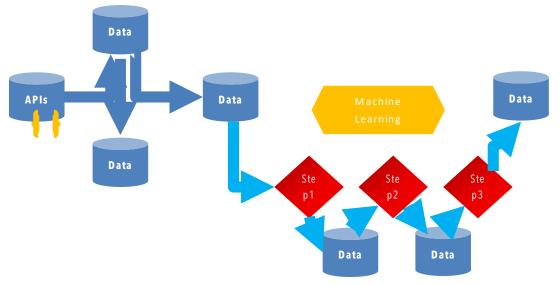
Signals, Logs, External Data, User Content

System of Record (Customers Bookings)

Transient Data (Cache, Session State)

Activity	Before	After
Data Engineering	1 Quarter	1 week
Machine Learning Experiments	1/month	1/day
Visibility by Customers	No	Yes
Granular Debugging	No	Yes

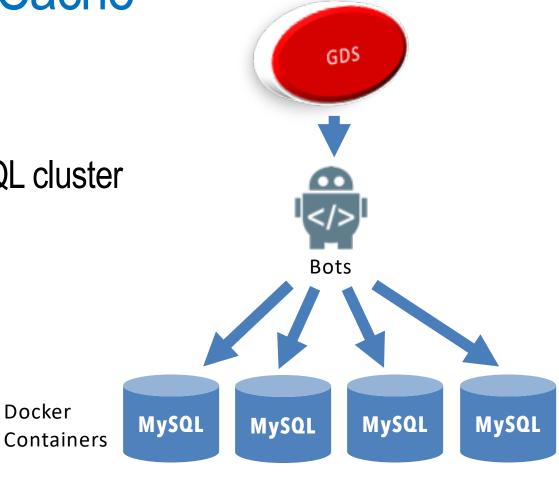
After ZIPLINE





After Machine Learning for Cache

- Cache hit ratio up from 61% to 96%
- Parity up from 82% to 98%
- No need to add more shards to MySQL cluster



Summary

- Machine Learning is useful for smart machine usage
- Use data for event based data collection and manipulation
- Think "pipe"; not ETL
- Data Engineering is the foundation for Data Science
- Airflow can be used to automate and quicken data engineering
- The python code can be generated easily by shell scripts
- Quicker to market means Data Scientists can experiment more frequently



Thank You!

Blog: arup.blogspot.com

Tweeter: @ArupNanda

Facebook.com/ArupKNanda