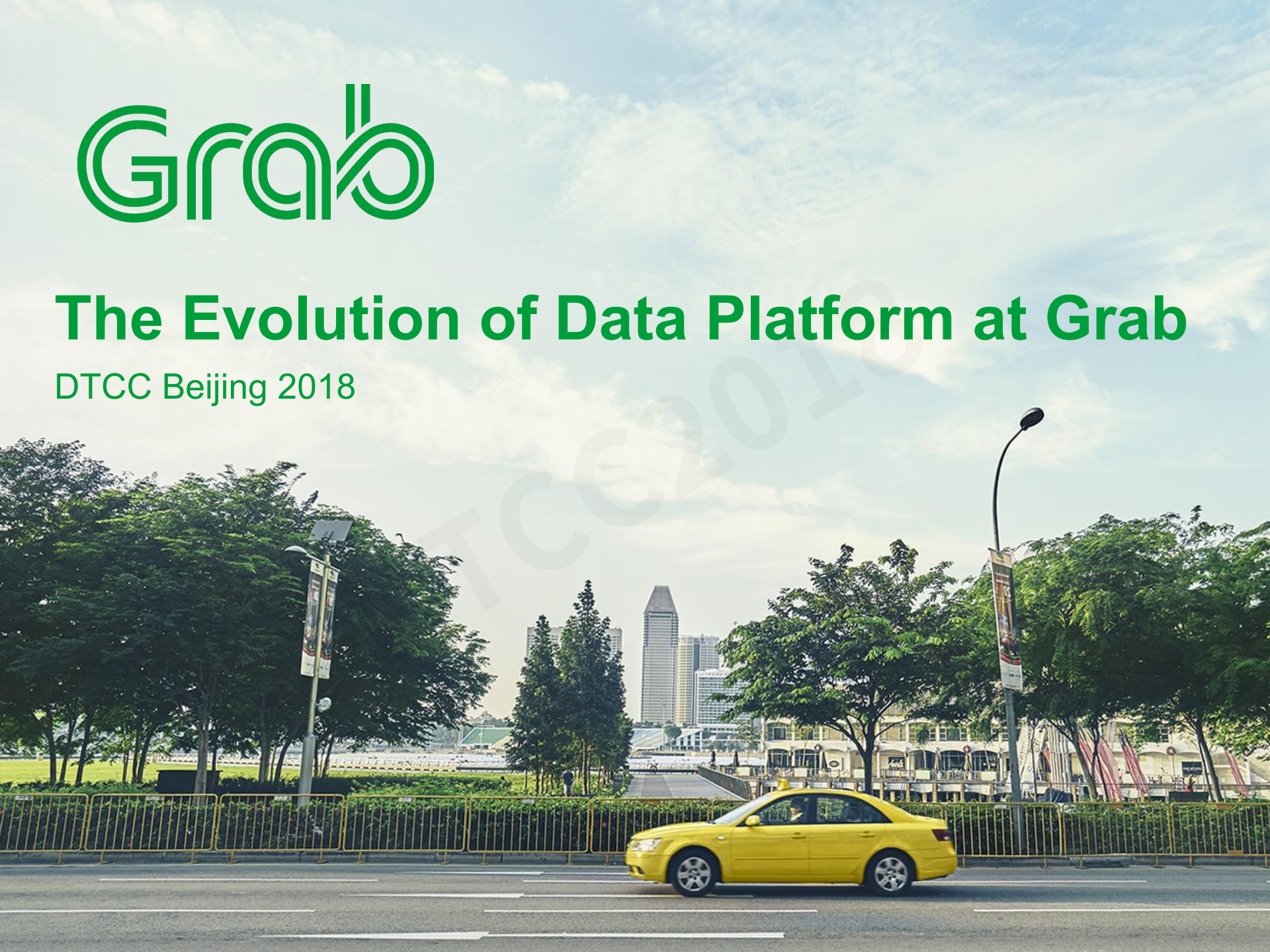
The Evolution of the Data Platform in Grab

Cheng Feng



The Evolution of Data Platform at Grab

DTCC Beijing 2018

#1 mobile app in SEA

1 Billion+
Rides annually

86 Million+
Downloads

2.6 Million+
Drivers in 191 cities

#1 transport app in SEA

Agenda

- **History, Data in Grab**
  - Data Analytics Platforms
  - Architecture review
- **Now, Why Presto, Lessons and Mistakes**
  - Decouple storage & compute
  - Stories and lessons learned
  - Data Gateway
- **Future, Serverless, Real-time platform**
  - Serverless data platform
  - Real-time platform

History, Data in Grab

Early Challenges

Architecture review

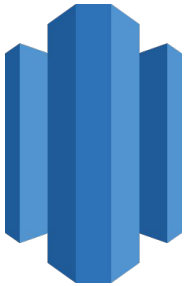
Early Challenges

- Hyper Local Strategy
  - Different markets track different metrics
  - Analytics is driven from the ground up
  - Hard to pre-process data due to these requirements

- Scaling to keep up with growth
  - Data Volume would double every 2 - 3 months
  - Number of reports, metrics and consumers growing.
  - Complexity of workloads also increase

Scalability and Concurrency Problems

- Simple stack, easy to manage for small team.



Data Warehouse Operational Reports Management Reports

- All workloads run on a massive Redshift Cluster
- Scalable but storage and compute are tightly coupled
- High concurrency and resource contention killed perf.

Decoupling Storage and Compute

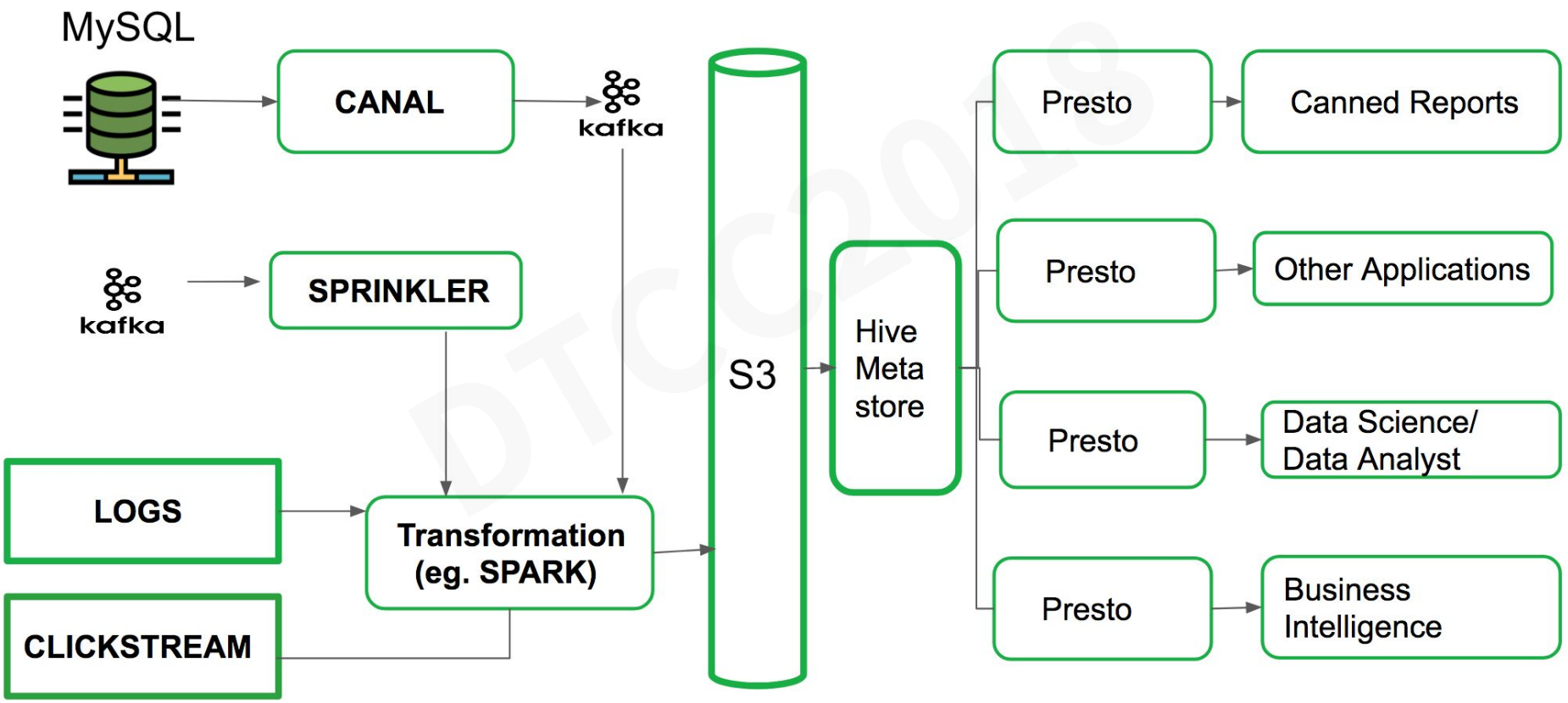
CLIENT

Data Gateway

COMPUTATION

STORAGE

Data Analytics Platforms



Key metrics

~2PB hot data
~15TB new daily

~20,000 jobs daily
~50,000 queries per day
in Presto

~20 clusters
~600 instances
~300 presto active users

Now, Why Presto, Lessons and Mistakes

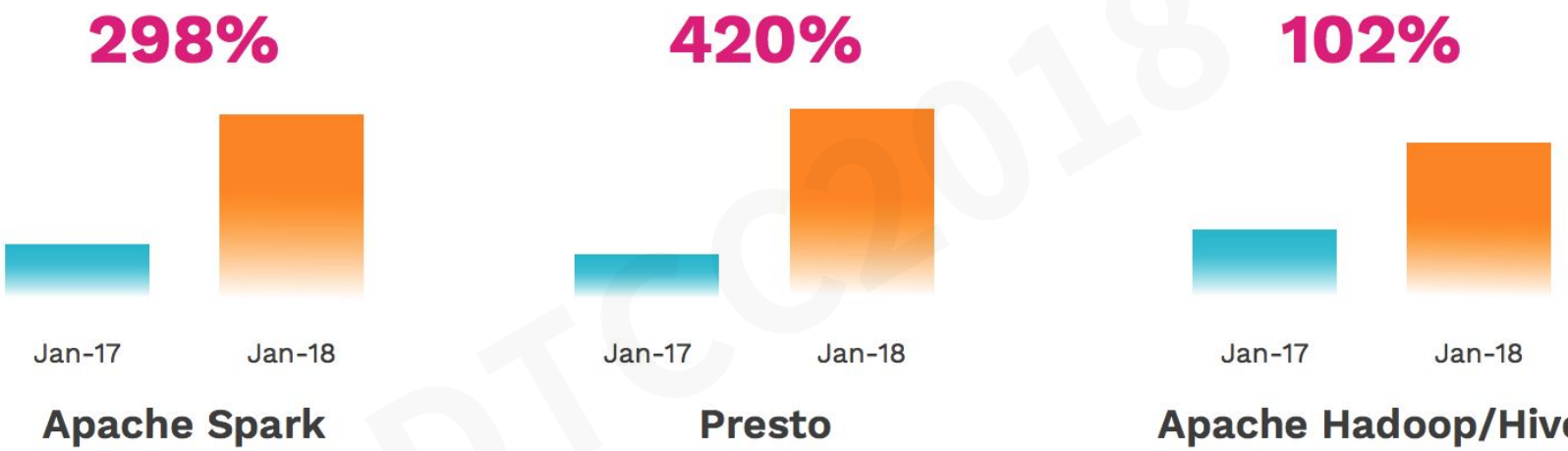
Decouple storage & compute

Presto, Stories and lessons learned

Data Gateway

BIG DATA ACTIVATION REPORT 2018

Total Engine Usage Globally by Compute Hours (YoY Change Jan '17–Jan '18)



Qubole®

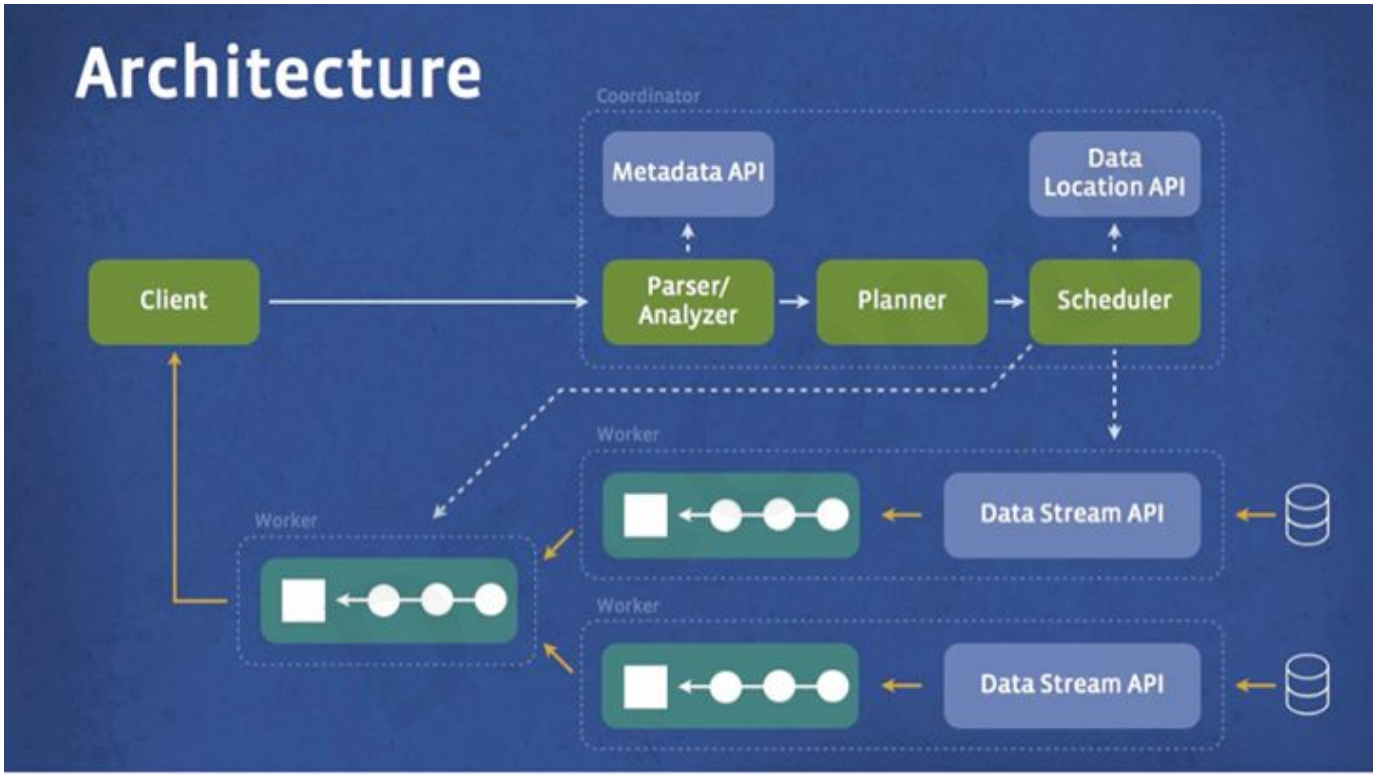
Presto

Understanding the memory management in presto to make trade-off between concurrency and memory Limit

Choosing the right EC2 Instance Type with correct configuration

Presto Performance Tuning

Presto Architecture



https://www.facebook.com/notes/facebook-engineering/presto-interacting-with-petabytes-of-data-at-facebook/10151786197628920/

Example

```
SELECT tab1.hour,
       tab1.pickup_geohash,
       Count(*) AS num
FROM   catalog.SCHEMA.bookings tab1
       INNER JOIN catalog.SCHEMA.cities tab2
               ON ( tab1.city_id = tab2.id )
WHERE  tab1.year = '2017'
       AND tab1.month = '11'
       AND tab1.day = '19'
       AND tab2.country_code = 'SG'
       ...
GROUP  BY 1, 2
HAVING Count(*) > 10000
```

PrestoCli

Presto JDBC Driver

Presto-cli

Data Gateway

Http

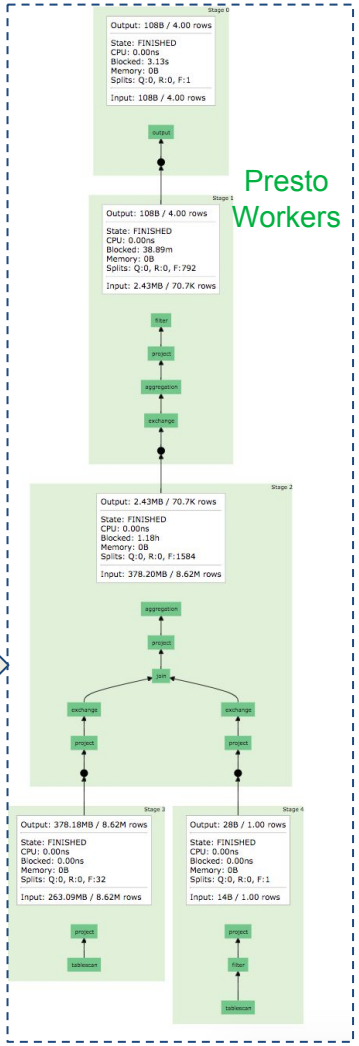
Query Results

PrestoCoodinator

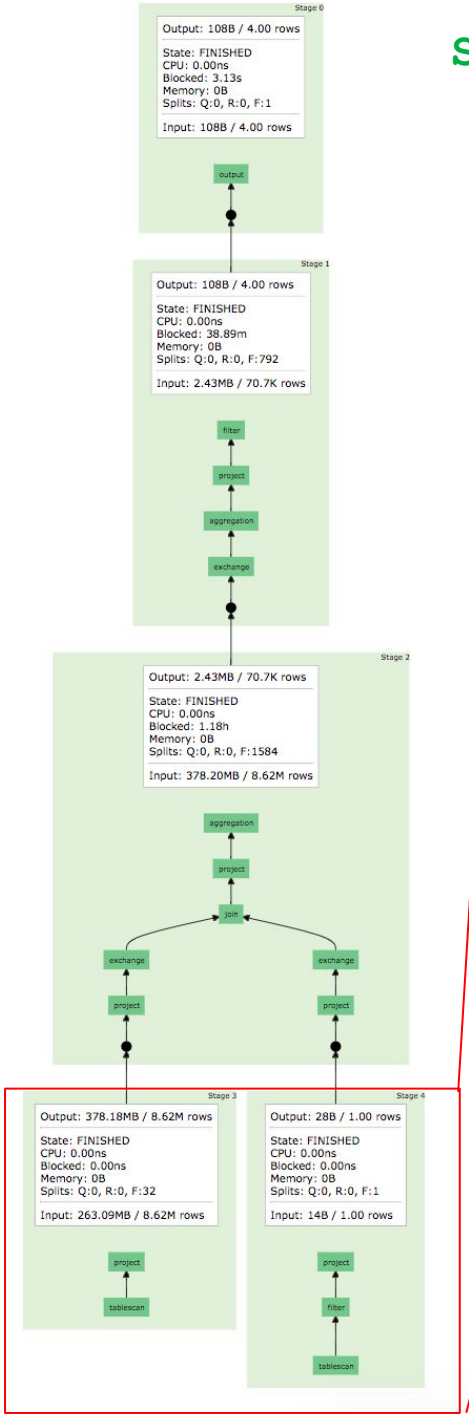
1. Parser the query and call metadata API

2. Come up with a logic plan, then physical plan with a series of stages

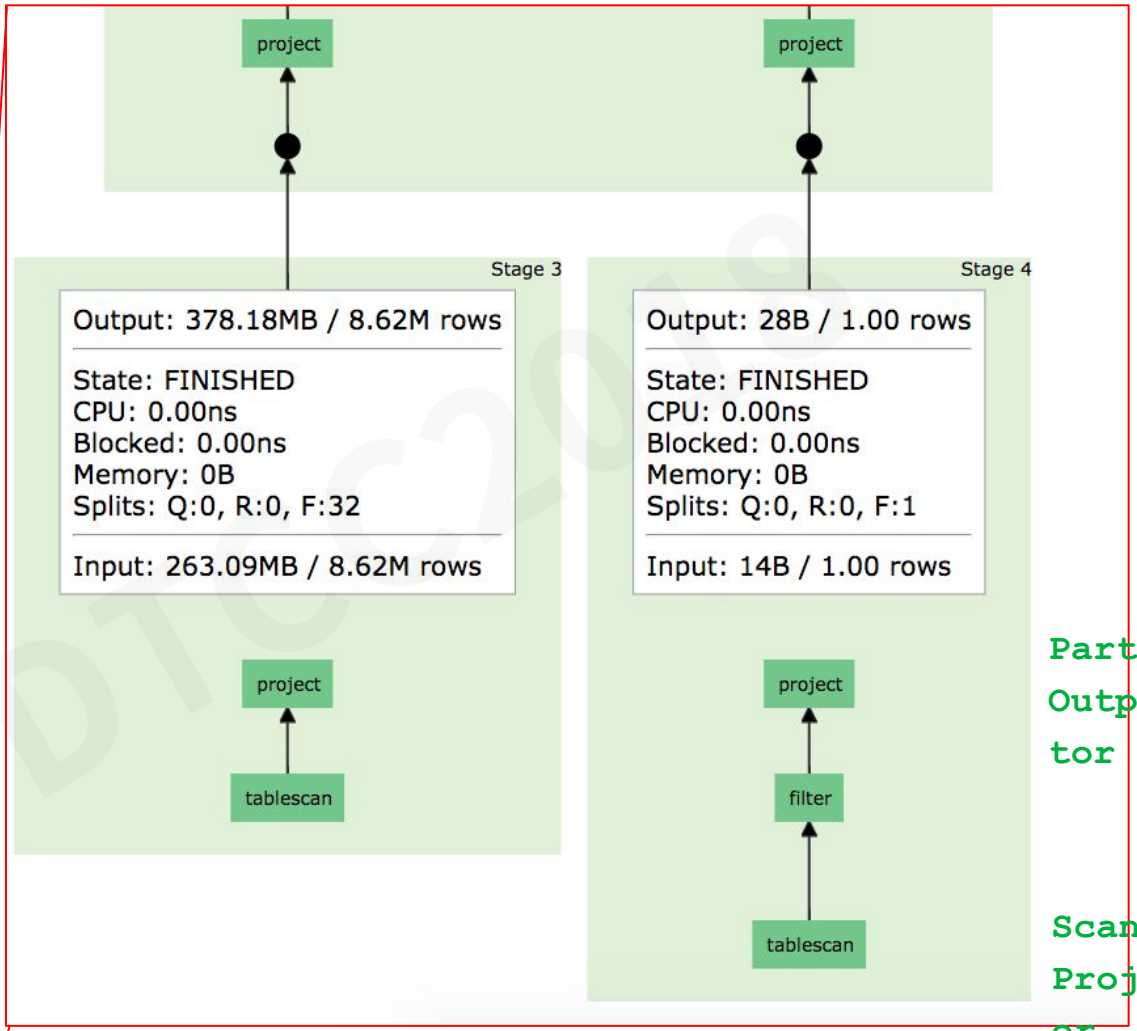
3. Stages will be translated into tasks running on a cluster of Presto workers



Presto Workers

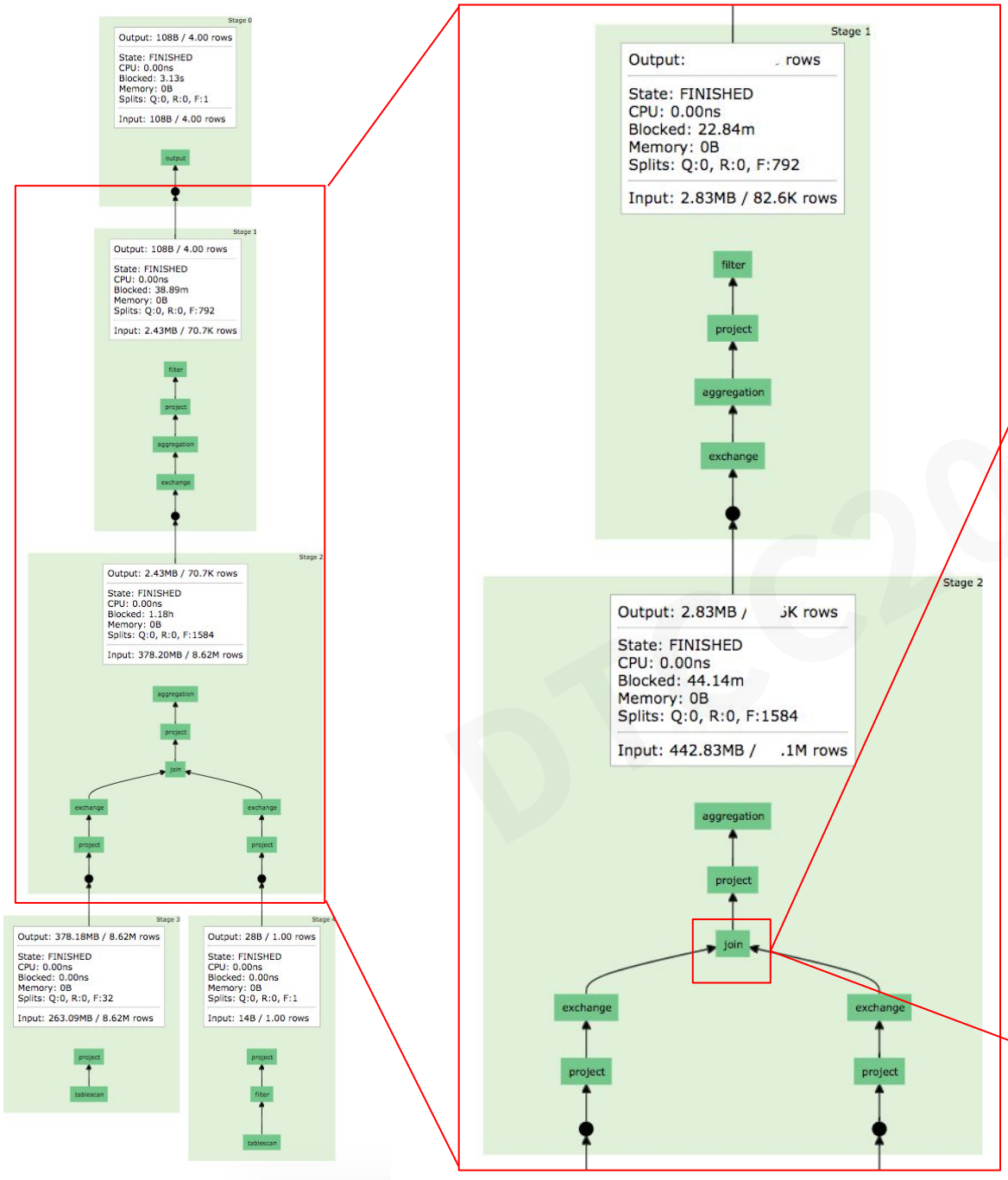


SOURCE stage



PartitionedOutputOperator

ScanFilterAndProjectOperator



HASH stage

Tasks operate on splits; Split is part of table/dataset;

Tasks contain one or more parallel drivers Drivers act a set of operators in memory

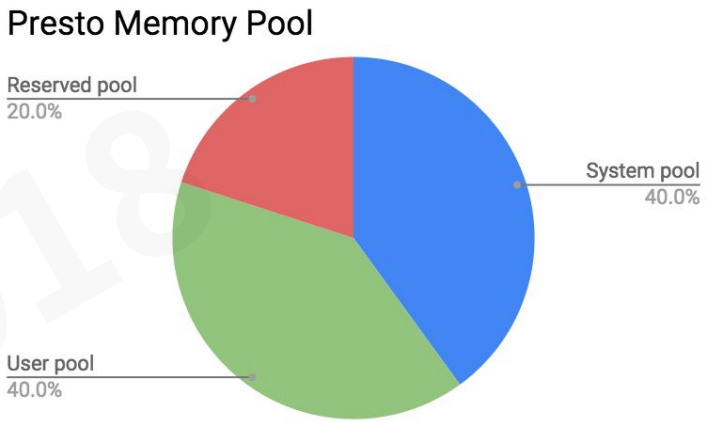
For example, A Join has HashBuilderOperator LookupJoinOperator

Error #1

Query exceeded max memory size of 30GB

Concurrency VS Memory Limit

- Presto Memory Pool
  - General pool(user+system)
    - Initially submitted to General Memory Pool
  - Reserved pool
    - Used only 1 time if GP exhausted
    - query.max-memory-per-node (Limit for per query per node)



Notes:

More Reserved Memory, support bigger query; (0% or 60%)

More General Memory (maxHeap-Reserved), support more concurrent Jobs.

Spilling memory to disk to avoid exceeding memory limits for the query.

https://github.com/prestodb/presto/issues/2624

https://github.com/prestodb/presto/issues/8638

Example

Case 1:

maxHeap=10GB,General pool=9GB; Reserved pool=1GB

If you have 8 queries need 1GB for each, and one need 6GB. The query(6GB) will fail, and we can have **8** queries running in general pool.

Case 2:

maxHeap=10GB,General pool=4GB; Reserved pool=6GB

If you have 8 queries need 1GB for each, and one need 6GB. The query need 6GB will be move to reserved pool, but in general pool, we can only run **4** queries parallely, and the other 4 queries will queuing until some queries finish execution.

Error #2

Encountered too many errors talking to a worker node. The node may have crashed or be under too much load.

Page Transport Timeout

- Infrastructure
  - Outbound network spikes hitting caps
  - Coordinator sending plan was costly
  - Workers saturating NICs
  - No fault tolerance

- Configuration
  - JVM size
  - GC type(Correlated GC pauses with errors => G1GC collector)
  - Tuned timeouts

Lesson #3

Presto Performance Tuning

Presto Optimization

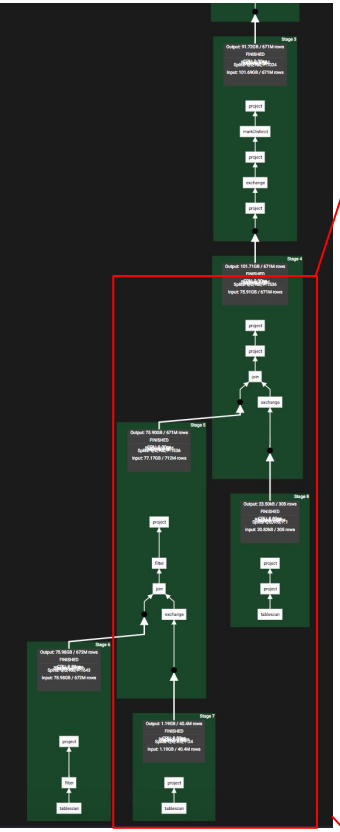
- Choosing the Right Storage Plan
  - Partition the data along natural query boundaries;
  - Use columnar data store;
  - Optimize file sizes (File Size, Compression).

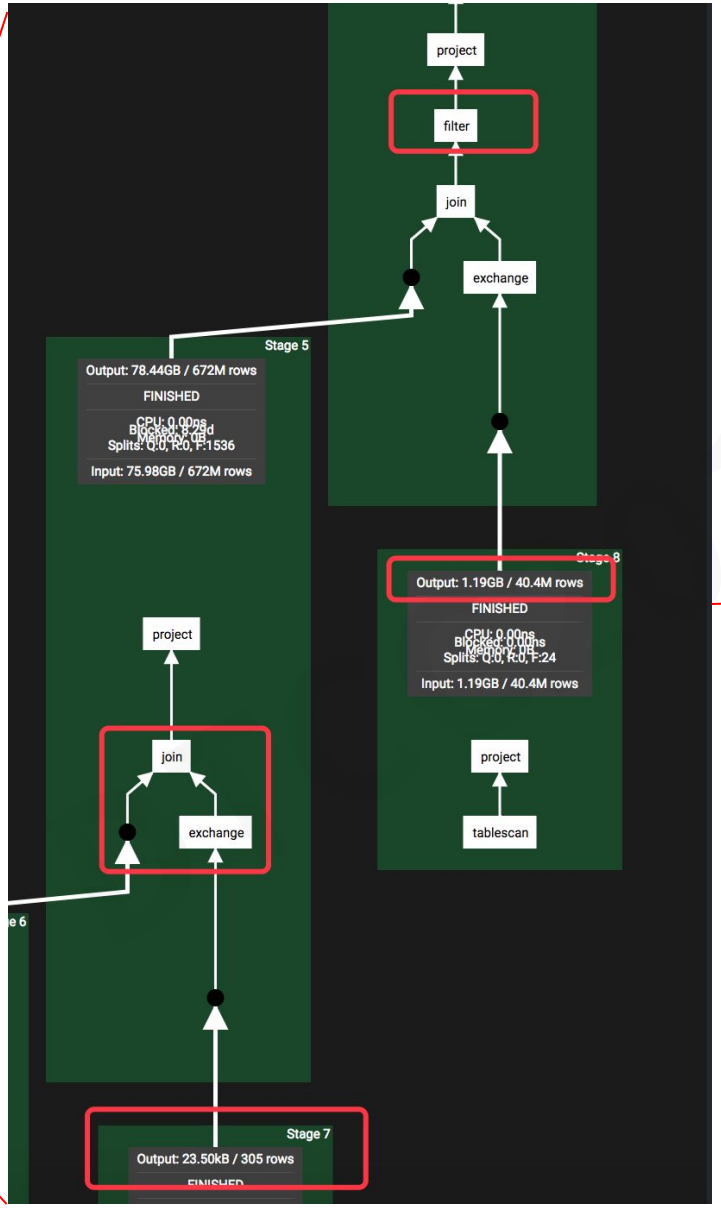
- Query optimization
  - Simple rule-based optimizer;
  - Reordering the join, for example: large tables first in join clause;
  - The cardinality within GROUP BY, for example: paxid,country;
  - Distributed join or broadcast join.

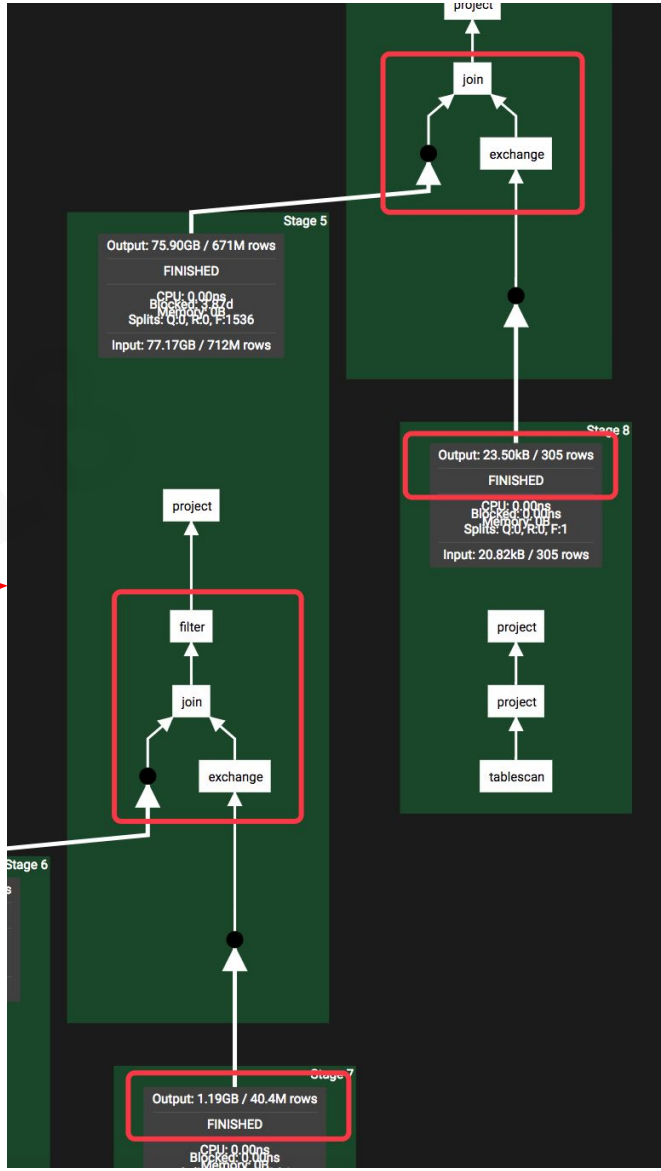
Case study

```
SELECT vertical,
       date_local,
       AVG(distance) AS avg_distance,
       COUNT(event) AS event_count,
       ...
FROM passenger_check_price tab1
  LEFT JOIN test_bookings tab2
        ON (tab1.bookingcode = tab2.code)
  LEFT JOIN taxi_types tab3
       ON (tab1.vehicletypeid = tab3.id
        AND tab1.streamtime >= tab3.start_at
        AND tab1.streamtime < tab3.end_at)
WHERE  concat(year,'-',month,'-',day,' ',hour,':00:00')
    >= date_format(now() -INTERVAL '30' day,'%Y-%m-%d')
AND tab2.code IS NULL
...
GROUP BY 1,
         2;
```

```
SELECT vertical,
       date_local,
       AVG(distance) AS avg_distance,
       COUNT(event) AS event_count,
       ...
FROM passenger_check_price tab1
  LEFT JOIN test_bookings tab2
        ON (tab1.bookingcode = tab2.code)
  LEFT JOIN taxi_types tab3
       ON (tab1.vehicletypeid = tab3.id )
WHERE tab2.code IS NULL
AND tab1.streamtime >= tab3.start_at
AND tab1.streamtime < tab3.end_at
AND  concat(year,'-',month,'-',day,' ',hour,':00:00')
    >= date_format(now() -INTERVAL '30' day,'%Y-%m-%d')
...
GROUP BY 1,
         2;
```



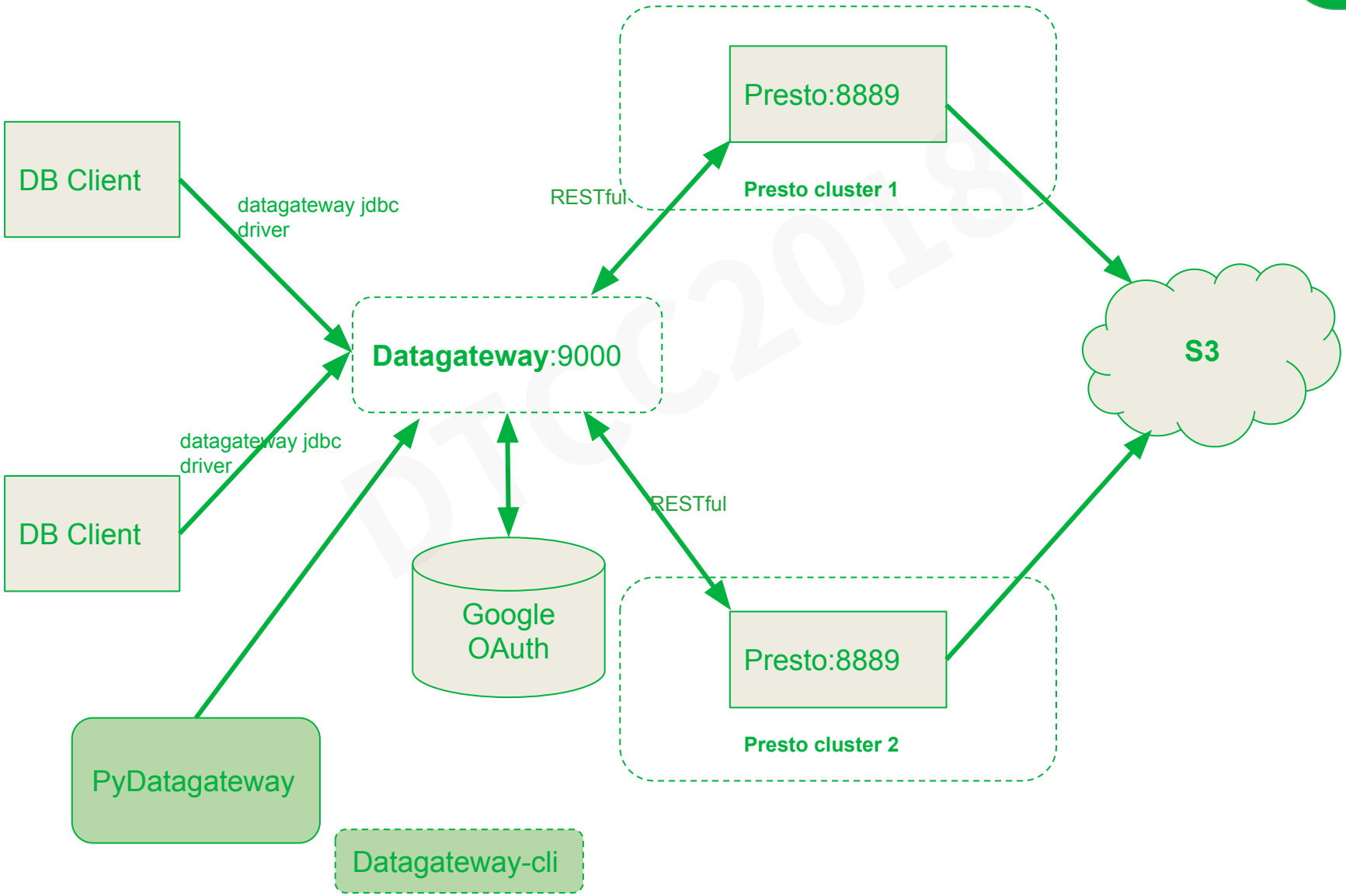




Before 13.79m

After 6.76m

Datagateway Design



Future, Serverless, Real-time platform

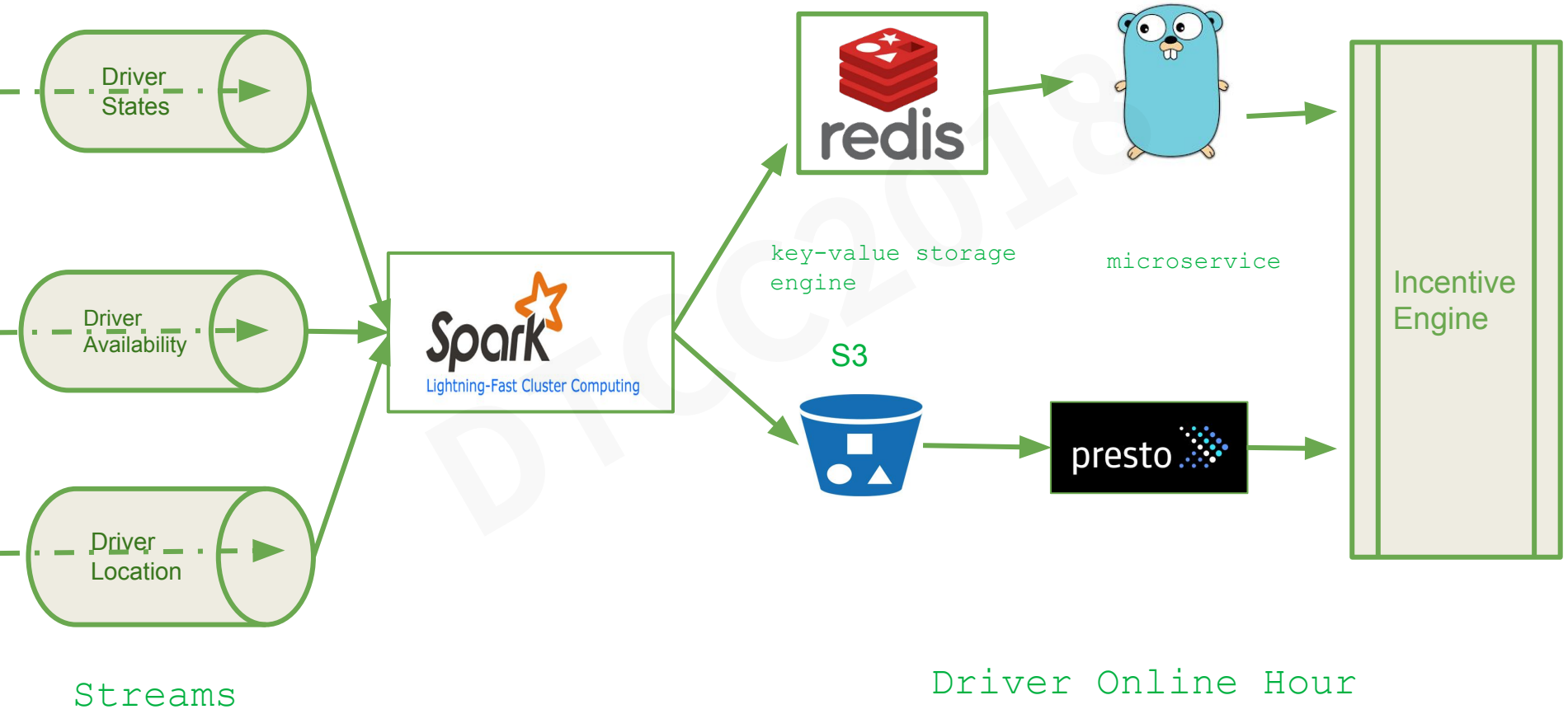
Serverless data platform

Real-time platform

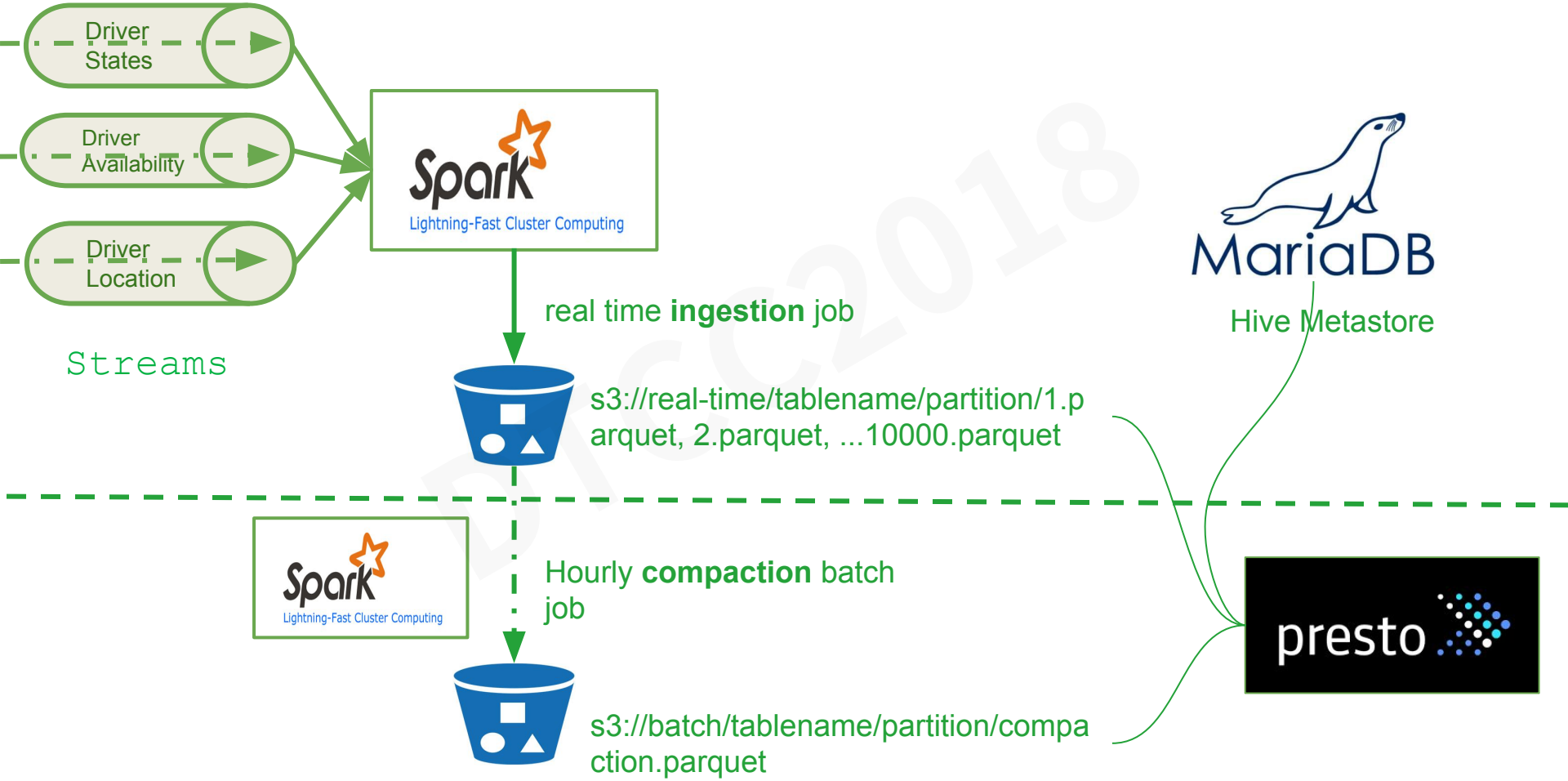
Moving to serverless data platform

- Athena
  - Serverless Presto compatible query engine;
  - No code changes migration, Cost saving (bill based on data volume scaned);
  -
- Glue
  - Serveless Spark compatible ETL service;
  - Automatically generates the code(billed by DPU)

Processing Data in real-time



Making Data available in real-time

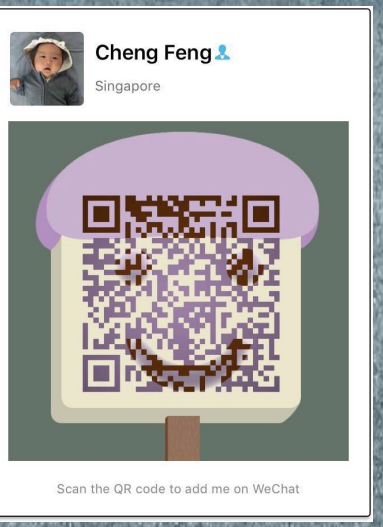


FORWARD, TOGETHER.

Forward, because we constantly aim to outserve the region by transforming inefficient systems for progress.

Together, because we work as one team to develop solutions that empower people and their livelihoods without silos.

Welcome to talk to me:
feng.cheng@grab.com



THANKS


HANA
Hadoop
SQL
MySQL
DB2
BigDate
Spark
SQL Server
NewSQL
Oracle
MPP
NoSQL
OLAP

讲师申请
联系电话（微信号）：18612470168

关注“ITPUB”更多
技术干货等你来拿~

与百度外卖、京东、魅族等先后合作系列分享活动

微学堂

让学习更简单

微学堂是以ChinaUnix、ITPUB所组建的微信群为载体，定期邀请嘉宾对热点话题、技术难题、新产品发布等进行移动端的在线直播活动。

截至目前，累计举办活动期数60+，参与人次40000+。

ITPUB学院

ITPUB学院是盛拓传媒IT168企业事业部（ITPUB）旗下
企业级在线学习咨询平台
历经18年技术社区平台发展
汇聚5000万技术用户
紧随企业一线IT技术需求
打造全方式技术培训与技术咨询服务
提供包括企业应用方案培训咨询（包括企业内训）
个人实战技能培训（包括认证培训）
在内的全方位IT技术培训咨询服务

ITPUB学院讲师均来自于企业
一些工程师、架构师、技术经理和CTO
大会演讲专家1800+
社区版主和博客专家500+

培训特色

无限次免费播放
随时随地在线观看
碎片化时间集中学习
聚焦知识点详细解读
讲师在线答疑
强大的技术人脉圈

八大课程体系

基础架构设计与建设
大数据平台
应用架构设计与开发
系统运维与数据库
传统企业数字化转型
人工智能
区块链
移动开发与SEO

联系我们

联系人：黄老师
电　话：010-59127187
邮　箱：edu@itpub.net
网　址：edu.itpub.net
培训微信号：18500940168