Dremio

MORE THAN A JOURNEY

滴滴

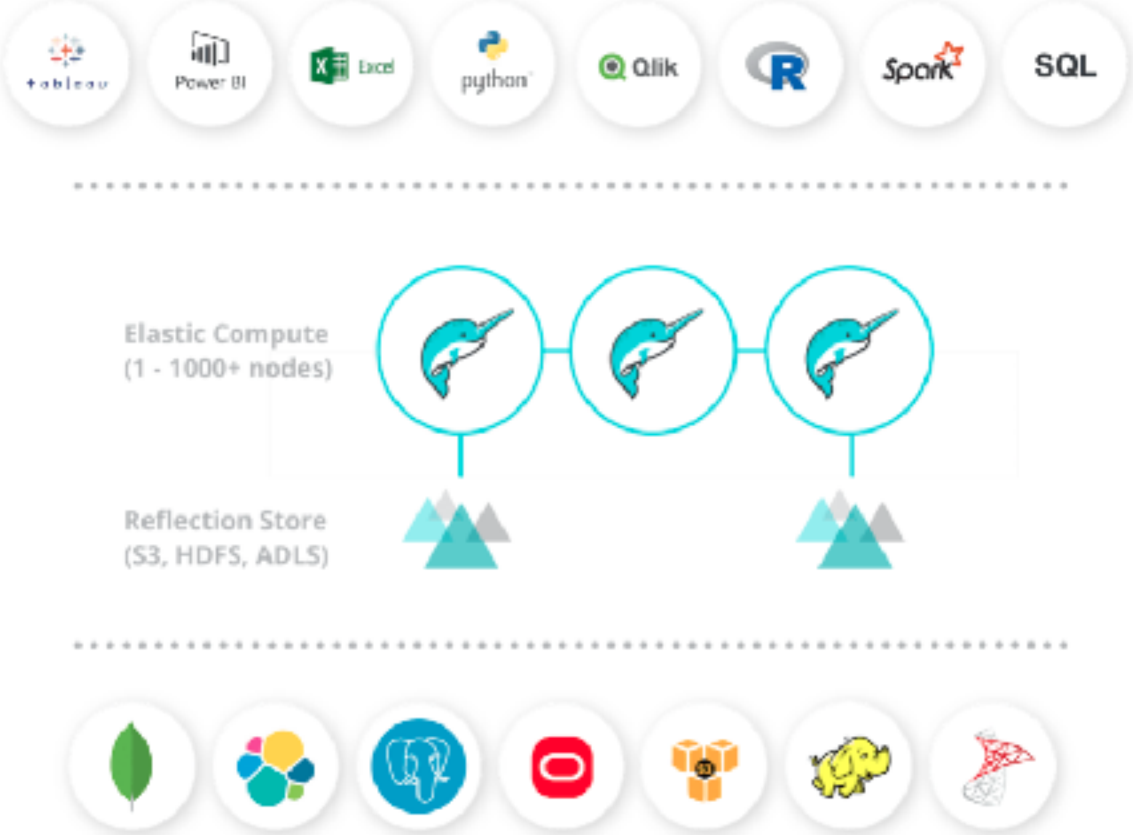
滴滴一下 美好出行

What

Dremio is the Data-as-a-Service Platform.

https://www.dremio.com/

Get more value from your data, faster. Dremio makes your data engineers more productive, and your data consumers more self-sufficient.



What

Apache Arrow Execution

From 1 to 1000+ nodes, architected for cloud deployments: elastic compute, runs on object stores.

Data Reflections™

Accelerate data and queries automatically, up to 1000x faster, with the full power of relational algebra.

Native Push-Downs

Optimized query semantics for each data source – Amazon S3, ADLS, RDBMS, NoSQL, HDFS, and more.

Vertically Integrated Query Engine

Cost-based query planner automatically generates query plans to make optimal use of Data Reflections™ and push downs.

MORE

Dremio = Apache Arrow + Sabot

- **Apache Arrow** https://arrow.apache.org/
  - Apache Arrow is a cross-platform standard for columnar data for in-memory processing. You can think of Arrow as **the in-memory counterpart to popular on-disk formats** like Apache Parquet and Apache ORC, and increasingly as the standard used by many different systems.
- **Sabot**
  - the engine inside Dremio
  - variant from drill （calcite & execution framework)

How

- **Data Source**
    - Hive Hdfs ES …
- **Data Set**
    - query result on data source
- **Reflections**
    - build on data set to accelerate data and queries
- **Query**
    - ansi sql
    - accelerate automatically, no need to change sql

Performance

- **Dremio**
  - NMG
  - 10 Workers, 1 Coordinator
- **Presto**
  - GZ
  - 14 Workers, 1 Master
- **Spark**
  - NMG
  - Yarn, max 100 executors

Performance

single table

```
select
    order_id, product_id, city_id, district, county,  starting_name,  dest_name, a_birth_time, strive_time
from hive.gulfstream_dwd.dwd_order_call_grab_d
where "year" = '2018' and "month" = '08' and "day" >= '01' and "day" <= '31'
```

SQL	Dremio	Dremio reflections	Presto	Spark
limit 100	2 s	2 s	2 s	5 s
count(order_id)	15 s	2 s	35 s	42 s
group by, cnt	10 s	3 s	90 s	150 s
group by, cnt, max avg	19 s	3 s	33 s	179 s

two table join

```
select count(1) from
(  select
     order_id, product_id, city_id, district, county, starting_name, dest_name, a_birth_time, strive_time
   from hive.gulfstream_dwd.dwd_order_call_grab_d
   where "year" = '2018' and "month" = '08' and "day" >= '01' and "day" <= '31'
) order_call_grab
inner join
(
   select order_id, driver_type
   from hive.gulfstream_dwd.dwd_order_make_d
   where "year" = '2018' and "month" = '08' and  "day" >= '01' and "day" <= '31'
) order_make_d
on order_call_grab.order_id = order_make_d.order_id
```

Performance

two table join

SQL	Dremio	left reflections	both reflections	Presto	Spark
count(order_call_grab.order_id)	29 s	26 s	17 s	72 s	88 s
group by, count	38 s	28 s	20 s	78 s	83 s
group by, count, avg	30 s	34 s	23 s	75 s	87 s

Performance

3 table join

```
select  count(order_call_grab.order_id) from
(  select
     order_id, product_id, city_id, district, county, starting_name, dest_name, a_birth_time, strive_time
   from hive.gulfstream_dwd.dwd_order_call_grab_d
   where "year" = '2018' and "month" = '08' and "day" >= '01' and "day" <= '31'
) order_call_grab
inner join
(
   select order_id, driver_type
   from hive.gulfstream_dwd.dwd_order_make_d
   where "year" = '2018' and "month" = '08' and  "day" >= '01' and "day" <= '31'
) order_make_d
on order_call_grab.order_id = order_make_d.order_id
inner join
(
    select
    order_id,
    city_id,
    driver_id
from hive.gulfstream_dwd.dwd_finance_order_target
where "year" = '2018' and "month" = '08' and "day" >= '01' and "day" <= '31'
) order_target
on order_call_grab.order_id = order_target.order_id
```

Performance

3 table join

SQL	Dremio	2 reflections	all reflections	Presto	Spark
count(order_call_grab.order_id)	43 s	31 s	29 s	79 s	124 s
group by, count	45 s	33 s	33 s	246 s	77 s
group by, count, avg	50 s	35 s	35 s	140 s	108 s

- **公司** hadoop hive 兼容性
  - 密码，federation
- **SQL 兼容性**
  - left join、outter join
  - ansi sql