VITESSE DATA

Data Warehouse in a PG / Greenplum / Vitesse DB Environment

CK Tan, Feng Tian
{cktan,ftian}@vitessedata.com

Intro

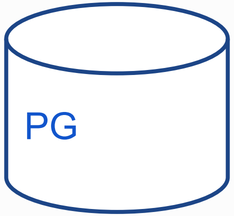
CK Tan

- Upwork, Greenplum, Informix

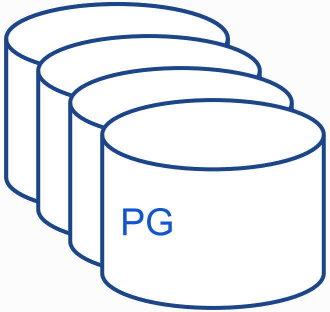
Feng Tian

- Datrium, VMware, Greenplum, Microsoft

In the beginning ...



Business was good ...



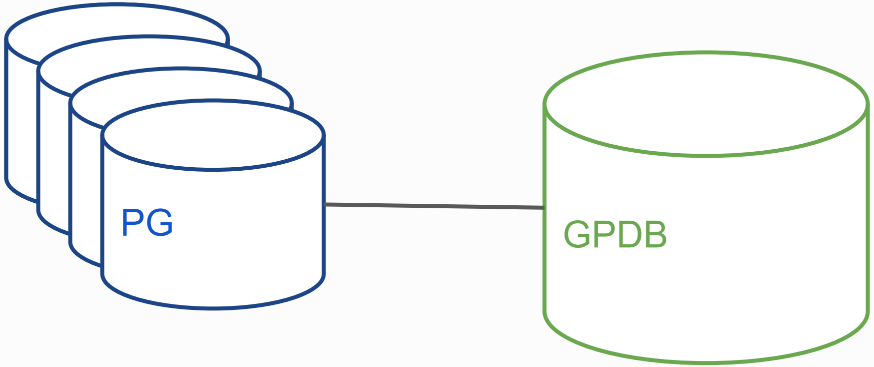
What is Data Warehouse?

OLTP	DW
Many single-tuple retrievals using indices	Big table seq scans
Queries run in 1ms	Queries runs for hours
Current data (ODS) < 1TB	Years of data - fact and dimension tables
100-1000 of connections	Handful of connections
Provides hot data to web users	Provides summary data to report users

Vocabularies

OLTP	DW
Indices	Window functions, Rollup, Cube
Referential Integrity	Table Partitions
TPS, TPC-C	ETL, ELT, MPP
	TPC-H, TPC-DS
Postgres, mysql, oracle, sybase, informix	Vertica, greenplum, exadata, teradata

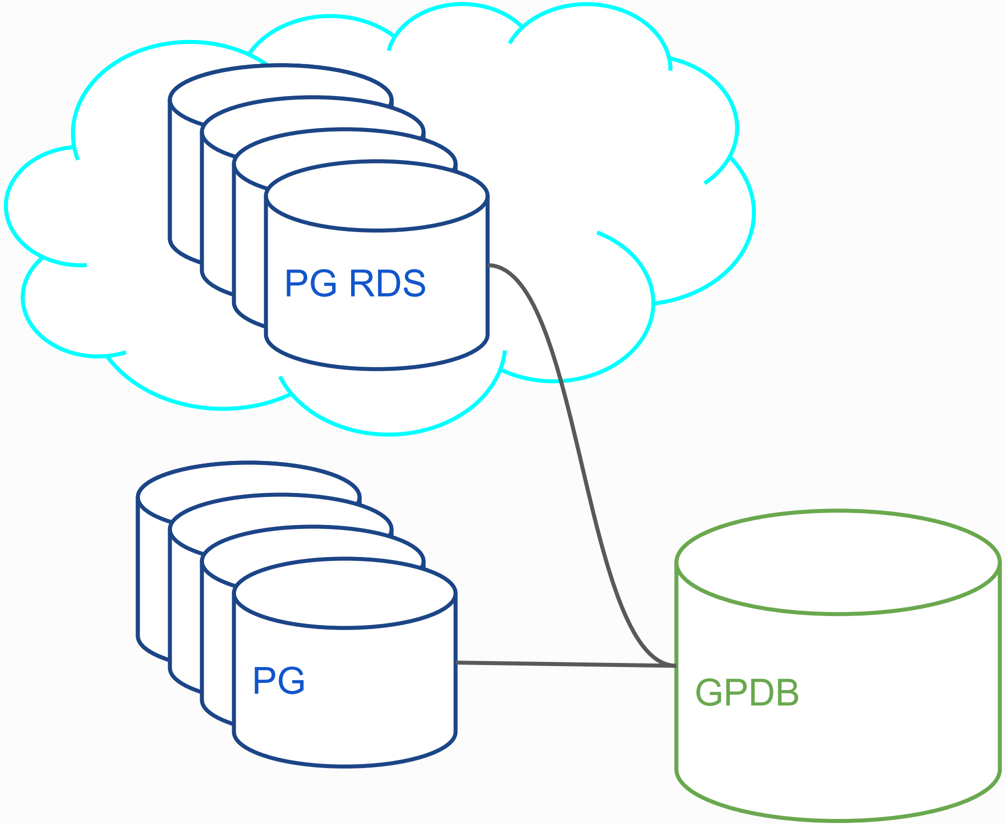
Analytics ...



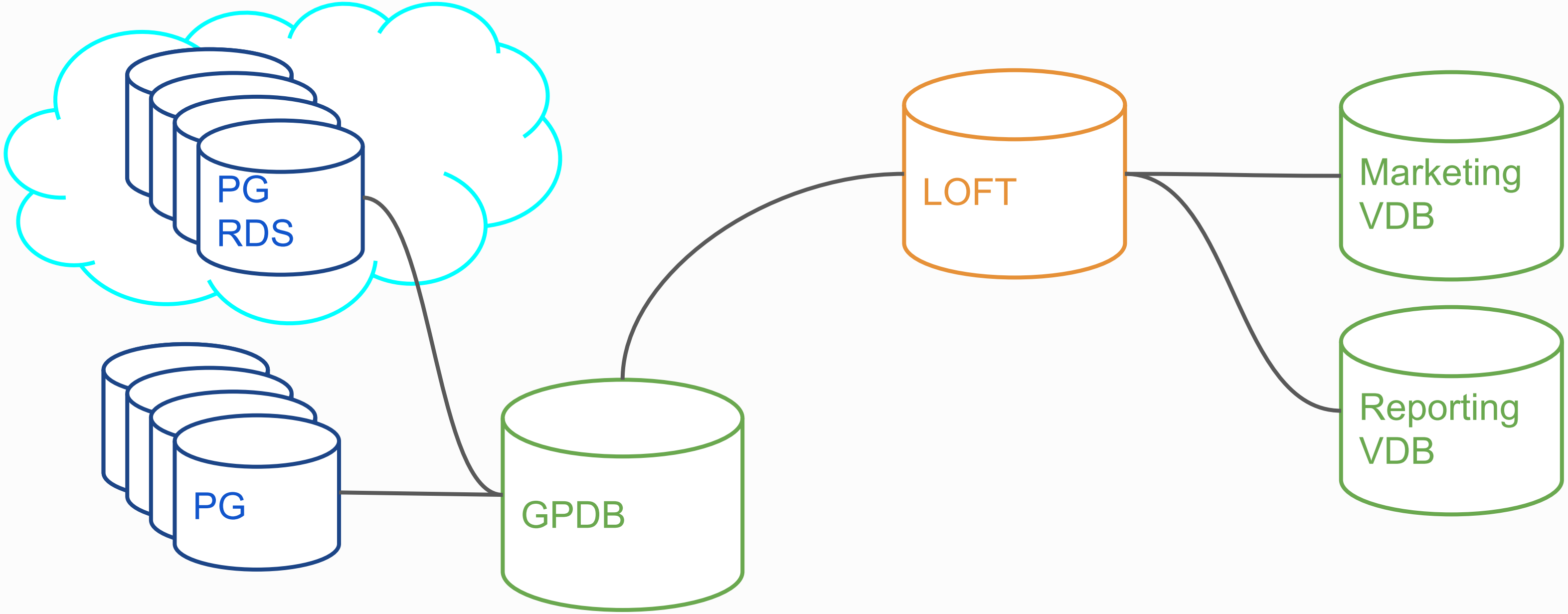
Appy ELT
* sync tables to GPDB
* apply aggregates

Transform
* apply aggs
* tag with interesting properties
* push agg tables into production

Microservices in AWS



Even More Analytics



Vitesse LOFT (Large Optimized Foreign Tables)

External column store

* File based
* Partition Aware
* SPQ: Simple Parquet Format

Query using Postgres Foreign Tables

Vitesse DB - PostgreSQL for DW

Inject new technologies

- JIT with LLVM
- Data-path optimization
- Column Store
- Threads

Scans go as fast as 18GB/s on heap tables

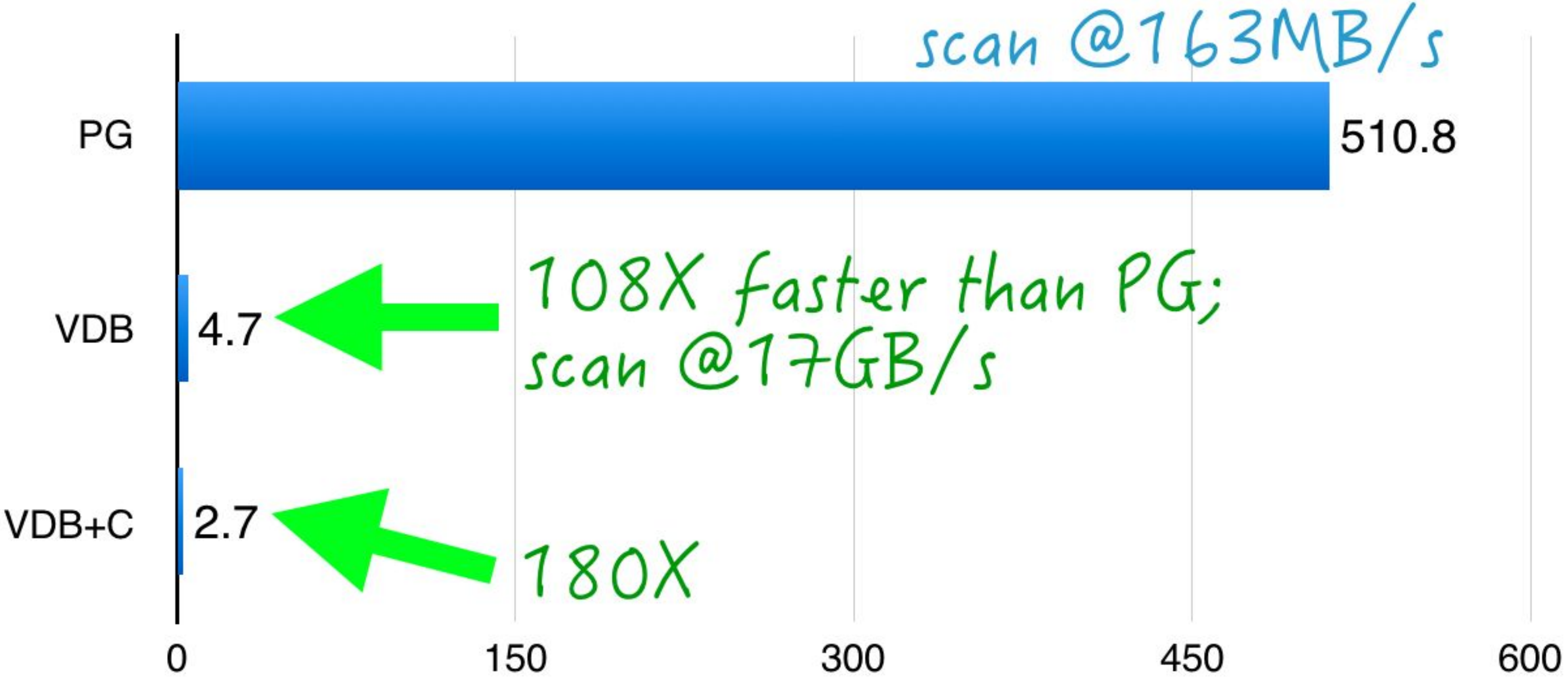
TPC-H 100, Q1 finishes in 3 sec

- PG takes 8.5 minutes

TPCH - Q1

```
SELECT l_returnflag, l_linestatus, SUM(l_quantity) as sum_qty,
    SUM(l_extendedprice) as sum_base_price,
    SUM(l_extendedprice * (1 - l_discount)) as sum_disc_price,
    SUM(l_extendedprice *
            (1 - l_discount) * (1 + l_tax)) as sum_charge,
    AVG(l_quantity) as avg_qty,
    AVG(l_extendedprice) as avg_price,
    AVG(l_discount) as avg_disc,
    COUNT(*) as count_order
FROM lineitem
WHERE l_shipdate <= date '1998-12-01' - interval '112 day'
GROUP BY 1, 2
ORDER BY 1, 2;
```

Q1 runtime in seconds (lower is better)

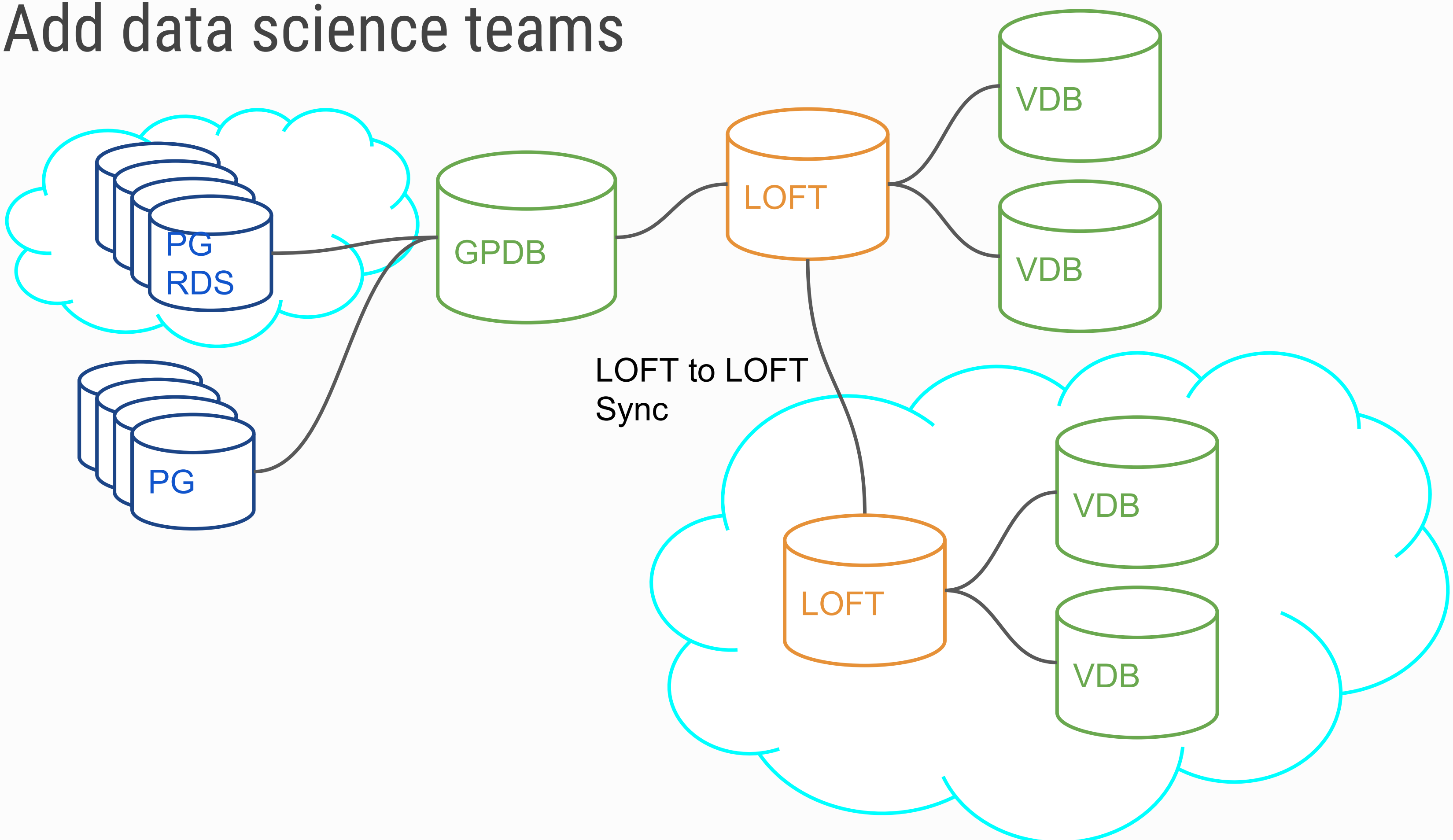


scan @163MB/s

108X faster than PG;
scan @17GB/s

180X

Add data science teams



Thank You

VITESSE DATA