

Greenplum new features and Roadmap

高小明

sgao@pivotal.io

Safe Harbor

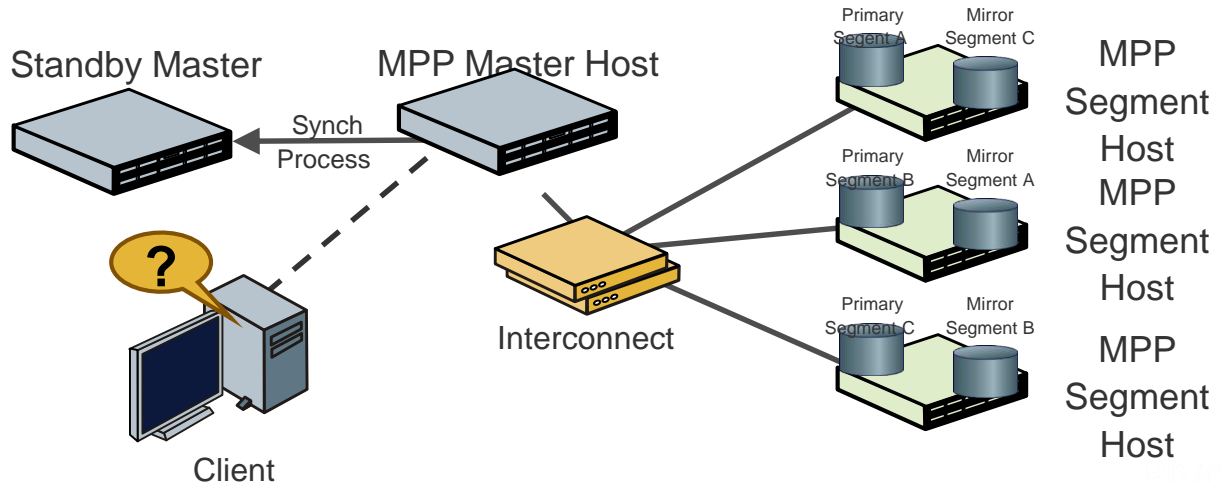
- *“Any information regarding pre-release of Pivotal offerings, future updates or other planned modifications is subject to ongoing evaluation by Pivotal and therefore **subject to change**. This information is provided without warranty of any kind, express or implied. Customers who purchase Pivotal offerings should make their **purchase decision** based upon features that are currently available. Pivotal has no obligation to update forward looking information in this presentation.”*

Agenda

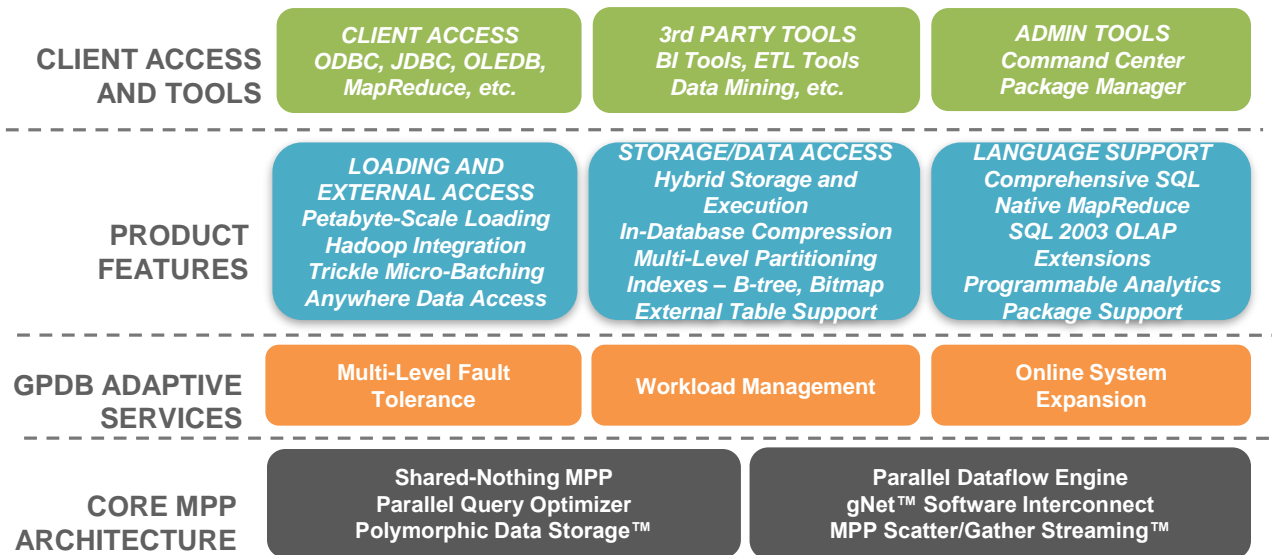
- Greenplum Overview
- New Features and Roadmap
- Airline Optimization Use Case

Pivotal

Greenplum Database Architecture



Key Features and Benefits of Pivotal Greenplum





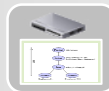
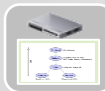
SQL



Master Servers

Query planning & dispatch

...



...

Network Interconnect



Segment Servers

Query processing & data storage

...



...

External Sources



cloudera

MAPR

Google Cloud Platform



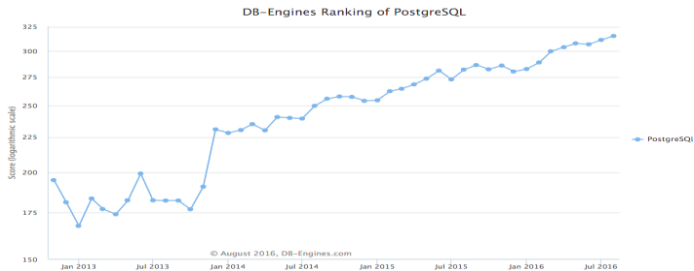
PostgreSQL Heritage



Greenplum
Open Source
Launch



PGCONF.EU 2015



- Widely used
- Open Source
- Enterprise class relational engine

Greenplum is Growing Steady

- Greenplum is Growing Steady
 - Operating in 34 countries globally
 - Customer count and revenue growing
 - Pivotal engineering investment growing
 - 9 Greenplum Database releases in 2016
 - Open source code contribution growing
 - 1417 commits to the github repo of Greenplum in 2016
 - 111 unique contributors on github repo of Greenplum in 2016
 - Major Greenplum 5.0 release planned early 2017

Pivotal

Greenplum Community

- <https://github.com/greenplum-db/gpdb>
- Since open source from 2015/10/27
 - GPDB: 1652 stars, 472 fork, 299 watch
- Contributions
 - Pull Request(PR): 31 open, 987 closed within 12 months
 - External contributions from China: China Mobile, Alibaba, Huawei, ...
- User groups without any advertisement
 - wechat group: 436



Agenda

- Greenplum Overview
- New Features and Roadmap
- Airline Optimization Use Case

Pivotal

PostgreSQL Base



PostgreSQL



Vision

Greenplum in the long run will be based on latest PostgreSQL

Upcoming Roadmap

- GPDB 5.0 release upgrade from PG 8.2 to PG 8.3 (2017 time frame)
- Refactoring to enable faster future PG upgrades
- JSON/JSONB
- Full Text Search
- Improved XML Type/Functions
- UUID Type
- Raster PostGIS
- Anonymous Code Blocks
- PostgreSQL based Analyze (faster)
- Extension Framework
- Foreign Data Wrapper (FDW)

Pivotal Query Optimizer - ORCA

- First Open Source Cost Based Optimizer for BIG data
- Applies broad set of optimization strategies at once
 - Considers many more plan alternatives
 - Optimizes a wider range of queries
 - Optimizes memory usage



TPC-DS 10TB, 16 nodes, 48 GB/node

Performance: Query Optimization

Vision

Our new cost-based optimizer, ORCA, will become the default optimizer in GPDB for all workloads, performing equal or better than legacy optimizer in all cases.

Current Status

Complex workloads for analytics produce large gains with ORCA

Upcoming Roadmap

- Parallelizing Union and Union All Queries
- Expanding ORCA's index support to larger class of predicates
- Reduce optimization time:
 - Auto-disable unnecessary transformations
 - Investigation: Optimization Levels

Performance: Query Execution

Vision

Dynamic Code Generation is a next gen performance enabling technology

New Features

- Asynchronous dispatcher to improve performance and scalability

Upcoming Roadmap

- LLVM Dynamic Code Generation for faster query execution
- More accurate query memory accounting internally
 - Optimizer and zlib memory usage accounting can be improved
- Reduce Intra-Transaction Memory
- Reduce Idle-Time Memory Usage
- Catalog data caching in the optimizer to speed short running queries

Resource Management

Vision

Build a stable, pluggable and scalable resource management for multi-user resource isolation.

Current Status

Collect requirements and spike on new design

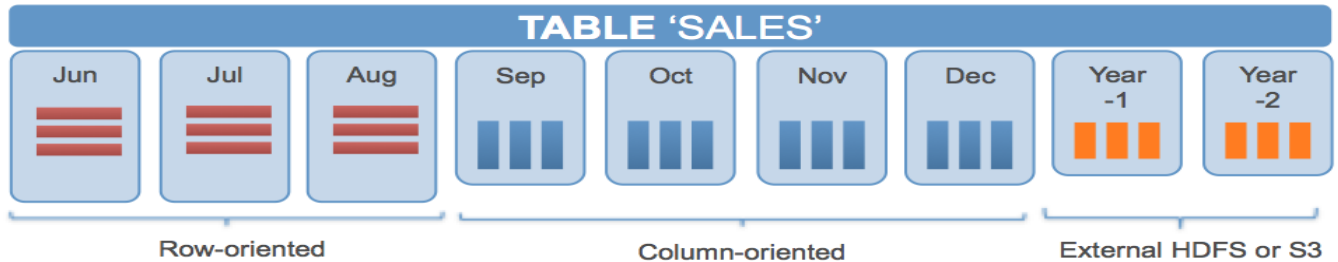
Upcoming Roadmap

- Disk-IO management
- Manage CPU by percentage
- Dynamically assign and change resource queue

Pivotal

Polymorphic Storage™

User Definable Storage Layout

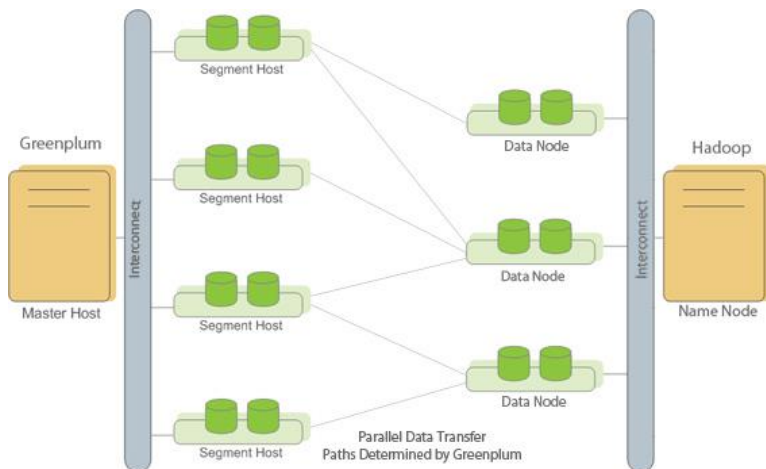


- Row oriented faster when returning all columns
- HEAP for many updates and deletes
- Use indexes for drill through queries

- Columnar storage compresses better
- Optimized for retrieving a subset of the columns when querying
- Compression can be set differently per column: gzip (1-9), quicklz, delta, RLE

- Less accessed partitions on external and seamlessly query all data
- All major Hadoop distributions
- Amazon S3 storage
- Others in development

GPHDFS: Greenplum HDFS Connector



Fully Parallel Data Transfers

External Table

Vision

- Wide variety of data source and targets for external data querying
- Leveraging external partitions on cheap and deep storage for online archiving

New Features

- S3 readable and writable External Tables
- GPHDFS enhancements for data type and HD distributions

Upcoming Roadmap

- Refactor and enhance GPHDFS with Pivotal libhdfs
- Support more cloud storages and data format

Pivotal

Backup & Restore

Vision

Increased support for mission critical systems

New Features

- Restore all views, indexes, user functions and sequence by schema
- Enhancement for data backup to Data domain and Netbackup
- Backup and restore support special characters in object names

Upcoming Roadmap

- PostgreSQL WAL Replication Segment Mirroring
- Reduce pg_class locking during backups
- Discovery investigations on next-gen backup improvements



Scalable, In-Database Machine Learning



- Recent updates: Pivoting, Path Functions, SVM Improvements
- Next areas of investigation: Pivotal R, Graph Analytics Functions, Misc.

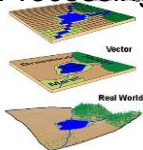
GPDB Geospatial



Current Key Features:

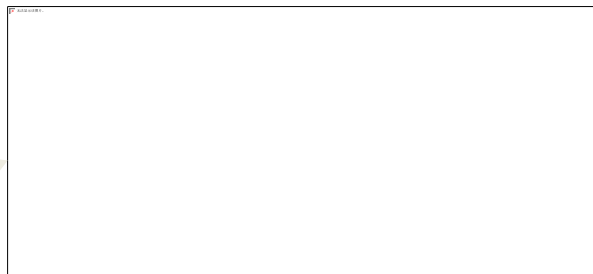
- Points, Lines, Polygons, Perimeter, Area, Intersection, Contains, Distance

Raster
Image
Processing



Ability to store
geospatial data
and query with
joins and
operators

Spatial Indexes &
Bounding Boxes



Round earth
calculations



Command Center (proprietary)

Vision

Clean and Rich Graphical User Interface for GPDB DBAs

New Features

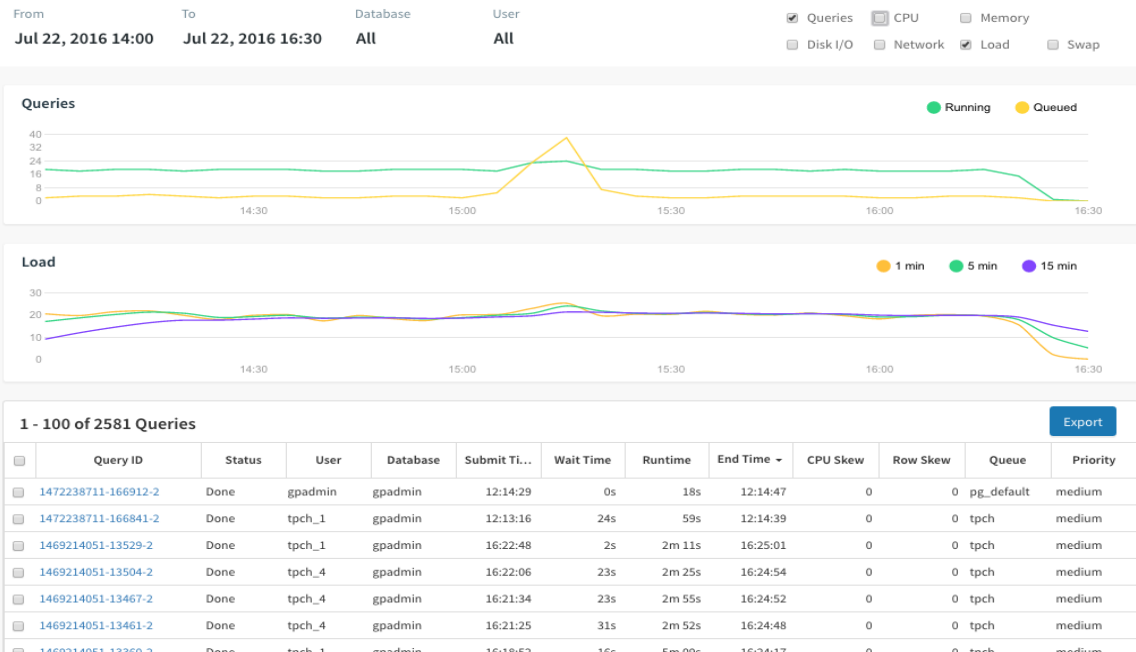
- HTML5 rewrite nearly at parity with original Flash-based GUI
- Richer dashboard and history

Upcoming Roadmap

- Kerberos integration + Security Theme
- Visual explain plan
- Integration with Greenplum Workload Manager for Graphical control of WLM

GPCC New UI Glance

Pivotal



Workload Manager (proprietary)

Rule based query management to monitor and manage queries and resource queues

New Features

- Include additional datum in actions
- Publish connections and sessions
- Improve response to cancel query

Upcoming Roadmap

- SUSE 11 support
- Improve spill file tracking
- User defined scripts as action

Pivotal

G2C (Greenplum Gemfire Connector) (proprietary)



Vision

Bring the real-time and high concurrency feature of Gemfire together with the full SQL analytics and reporting of Greenplum into an “Operational Data Warehouse” solution that combines the benefits of both

New Features

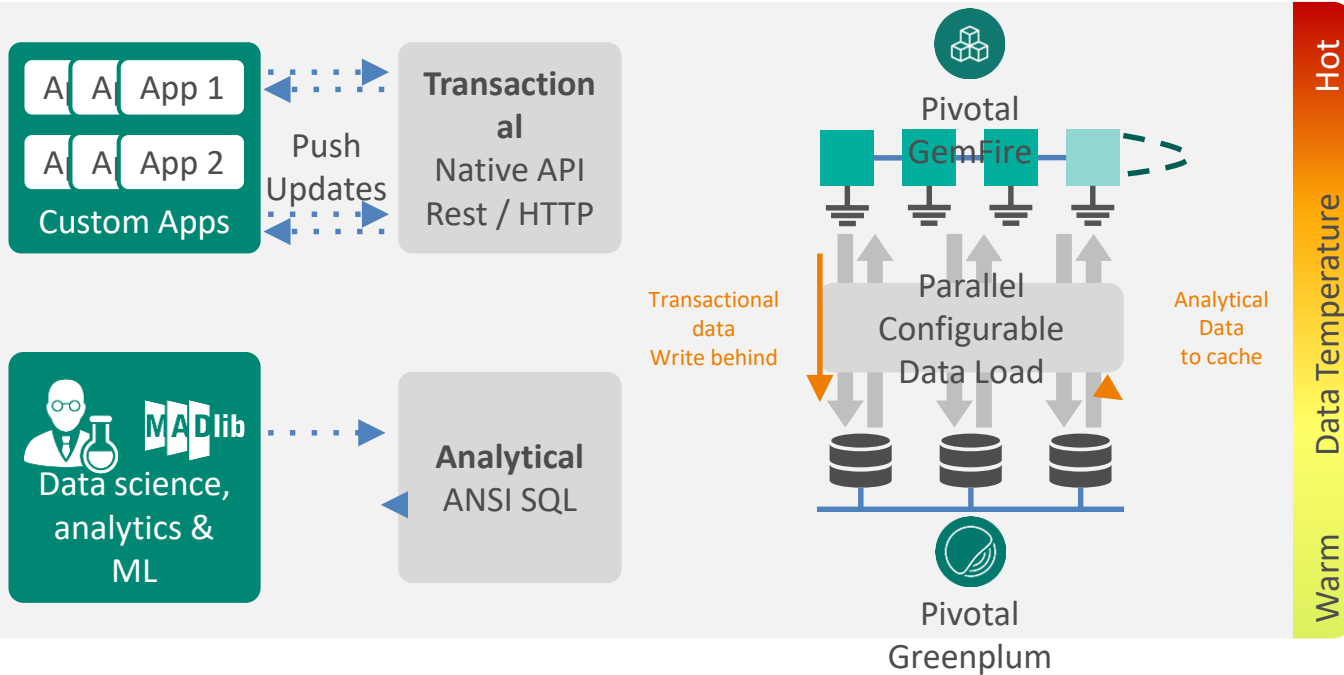
- IMPORT/EXPORT operations to/from Gemfire and GPDB

Upcoming Roadmap

- Greenplum based External Tables to provide READ/WRITE to Gemfire

Pivotal

GemFire and GPDB - Big Data Meets Fast Data



GPDB on Cloud

Vision

Bring GPDB to multi cloud platforms and Pivotal Cloud Foundry ecosystem with a smooth deploy and provisioning experience.

Current Status

- GPDB available in Azure Marketplace
- Beta release of BOSH deployment for GPDB

Upcoming Roadmap

- Easily deploy GPDB in more public Cloud platforms
- Incrementally improve Day 2 operations

Pivotal

PL/Container (proprietary)

Vision

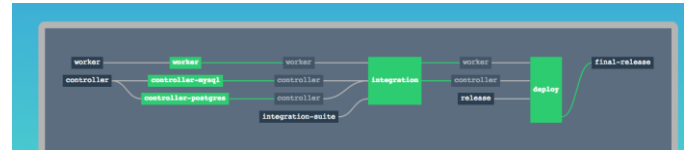
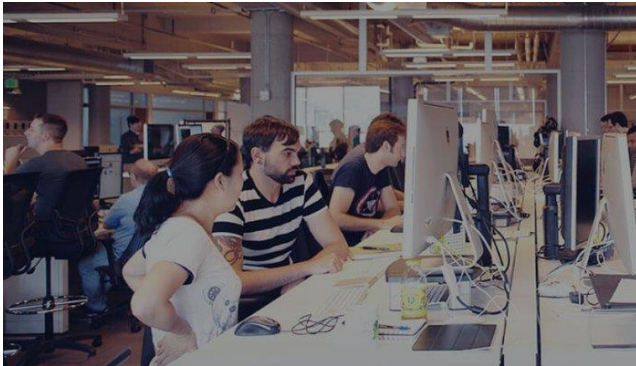
Containerized execution of Python and R (PL/Python and PL/R) providing a security model and an isolated environment to install the interpreter and any dependent libraries independent of the Database and DBA environment

Upcoming Roadmap

- Docker based containers
- Features which improve usability

Pivotal

Pivotal Engineering Practices



- Development teams in multiple geos
- Co-located with Pivotal Global Support Services
- Same methodology as all Pivotal teams

Agenda

- Greenplum Overview
- New Features and Roadmap
- Airline Optimization Use Case







Pivotal

Revenue Management

- What is Revenue Management ?
 - Charge different prices to different segments to maximize revenue
- When is it used?
 - There is a fixed amount of resources available for sale
 - The resources to sell are perishable
 - Customers are willing to pay a different price for using the same resources.
- What is the current state of Revenue Management?
 - Revenue Management is in use since early 1980s
 - Technology is dispersed
 - Competitor prices are now available
- What we will demonstrate
 - We wanted to combine machine learning and optimization in the same flow to highlight the benefits of having a single address for analytics.
 - Highlight two technologies: MADlib and PL/R

Data Generation - Airports

- We imported airport information with a Web Query from Wikipedia

City 	FAA 	IATA 	ICAO 	Airport 	Role	Enplanements 
ALABAMA						
Birmingham	BHM	BHM	KBHM	Birmingham-Shuttlesworth International Airport	P-S	1,443,215
Dothan	DHN	DHN	KDHN	Dothan Regional Airport	P-N	41,453
Huntsville	HSV	HSV	KHSV	Huntsville International Airport (Carl T. Jones Field)	P-S	606,127
Mobile	MOB	MOB	KMOB	Mobile Regional Airport	P-N	277,232
Montgomery	MGM	MGM	KMGH	Montgomery Regional Airport (Dannelly Field)	P-N	194,540
ALASKA						
Anchorage	ANC	ANC	PANC	Ted Stevens Anchorage International Airport	P-M	2,599,313
Anchorage	MRI	MRI	PAMR	Merrill Field	P-N	15,206
Anchorage	LHD		PALH	Lake Hood Seaplane Base (also see Lake Hood Airstrip)	P-N	15,184
Aniak	ANI	ANI	PANI	Aniak Airport	P-N	18,526
Barrow	BRW	BRW	PABR	Wiley Post–Willi Rogers Memorial Airport	P-N	40,674
Bethel	BET	BET	PABE	Bethel Airport	P-N	140,291

Route Creation: Hub & Spoke Model

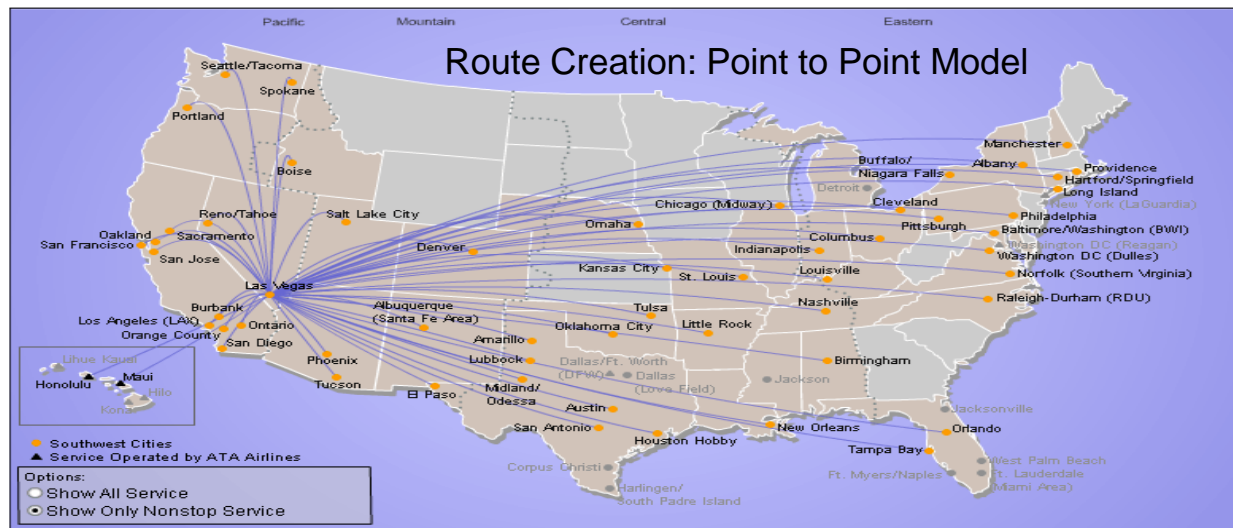
Routes arranged by Airlinerouteup.com
Original map copyright American Airlines

Routes arranged by Airlineroutemaps.com
Original map copyright **American Airlines**

Route Creation: Hub & Spoke Mode

Routes arranged by Airlineroutemaps.com
 Original map courtesy: AmericanAirlines.com

Routes arranged by Airlineroutemaps.com
Original map copyright **American Airlines**



Southwest Airlines route network maps from key focus cities

- | | | | | | |
|-------------------------------|------------------------------------|--------------------------------------|-----------------------------------|-------------------------------------|-------------------------------|
| Q Phoenix, AZ | Q Las Vegas, NV | Q Chicago-Midway, IL | Q Baltimore, MD | Q Houston-Hobby, TX | |
| Q Oakland, CA | Q Los Angeles, CA | Q Dallas-Love, TX | Q San Diego, CA | Q Nashville, TN | Q Orlando, FL |
| Q Tampa, FL | Q Philadelphia, PA | Q Albuquerque, NM | Q Kansas City, MO | Q Austin, TX | |



Data Generation

Airports:

- Serve only airports with annual enplanement larger than 1 M
- 65 Airports

Routes:

- Picked the Hub & Spoke Model
- From any airport to any other airport through one of the 5 hubs
- From/to any airport to/from any hubs non-stop routes
- No two connection routes allowed!
- 4180 Routes (640 are nonstop others through a hub)

Daily Flights:

- Up to 5 times for a flight
- 2200 flights/day

How realistic is 2200 flights/day?

- According to Bureau of Transportation Statistics the average number of flights per day in 2011 by Airlines

- American Airlines ~1,500
- Delta ~2000
- JetBlue ~ 600
- SouthWest ~ 3200
- United ~ 910

Summary Statistics Airline

Airline: American Airlines (AA)

Time Period: June 15, 2011 to June 15, 2011

Note: A complete listing of [airline](#) and [airport](#) abbreviations is available. Times are reported in loc

[Excel](#) | [CSV](#)

Carriers	All Flights					
	Total Number	Average Departure Delay (minutes)	Average Taxi-Out Time (minutes)	Average Scheduled Departure to Take-off (minutes)	Average Arrival Delay (minutes)	Average Airborne Time (minutes)
AA	1,546	11.26	16.11	27.37	8.67	141.88

- Note: Numbers presented here are averages. There are small seasonal deviations.

Data Generation

- **Sales History:**

- Close to two years of history
- Own Price + Competitors Prices
- Flight Date, Month, Weekday, Holiday Indicators
- Flights are available starting only 20 days before the flight date
- Sales for each sales date is captured.
- 3 Classes: First Class, Business, and Economy

4180 Routes	X	665 Flight Dates	X	Up to 5 Flight Times	X	20 Sales Dates	X	3 Classes
----------------	---	------------------------	---	----------------------------	---	----------------------	---	--------------

> 500 Million observations (> 150 GB)

Not big data but big enough that will not fit in memory

Note: With networks data grows exponentially! Remember we started with only 65 airports.

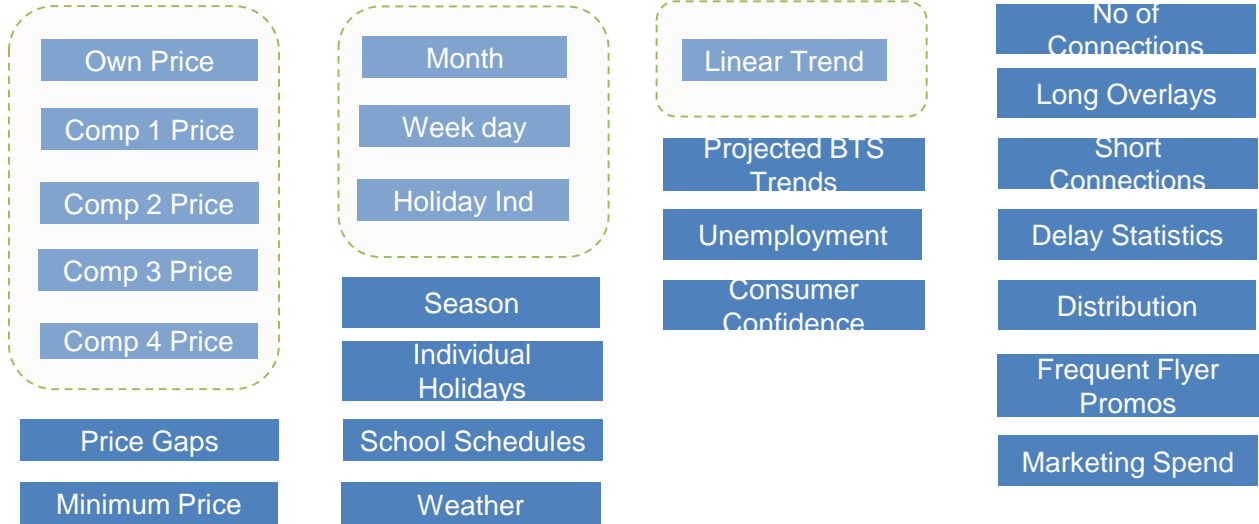
Problem Statement

- Decision Variables
 - p_t where t is Number of Days to Flight
- Assume linear relationship between Demand and Price
 - $D_t(p_t)$ is demand on day t when price is set to p_t
- Demand depends also on
 - Competitor Prices
 - Trend
 - Seasonality
 - Day of the Week, Month, Holiday Indicator

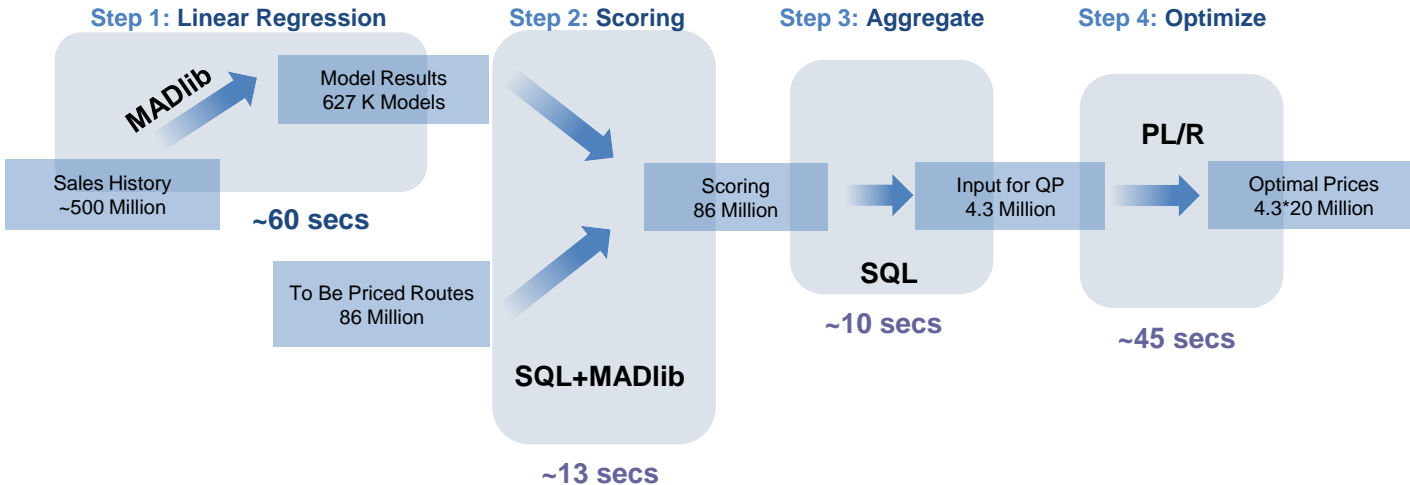
Once you determine $D(p_i)$ s - $(a_i p_i + b_i)$ the problem is a Quadratic Programming with Linear Constraints

$$\text{Max } \sum_i D(p_i) p_i \text{ st. } \sum_i D(p_i) \leq \text{Capacity}, D(p_i) \geq 0$$

Linear Regression: Other possible features



Solution



- Get insight from sales history
- Optimize the pricing decisions for 4.3 Million flights



- Approximately 2 Minutes
- With 4 Select Statements
- Without data ever leaving the DB

Thanks!

Q & A