Welcome
Advancing your Big Data Strategy

Robbin Cottiss
Strategic Customer Consultant
Tableau

Vindy Krishnan
Senior Product Manager
Tableau
You Know Me...
And Me…
Audience Poll

How many of you are new to Tableau?
How many of you are new to Big Data?
How many of you have or are moving their Big Data Platform to the cloud?
Our Goals for Today

Re-introduce “Big Data” and why it might be just “Data”

Understand how the Tableau Platform Integrates with Modern Data Platforms

Extract some patterns and generalize to provide a reference framework

Evaluate successful customer strategies and architectures

Demo the framework and recap

The Goal of Goals: Develop a Platform Mindset
What’s the Situation?
Converging Waves

Data

Cloud
Converging Waves
Data
Cloud
Modern Analytics Platform
Tableau
Why Do We Need a Modern Data Architecture?
The Tableau Platform

1. **Data Access**
   - Live | In-memory | Hybrid Connectivity

2. **Data Prep**
   - Data Blending | Query Federation | Visual Data Prep | Auto Data Modeling

3. **Governance**
   - Centralized Data Sources | Certification | Usage Analysis | Permissions

4. **Content Discovery**
   - Projects | Recommendations | Versioning | Search

5. **Analytics**
   - Visual | Ad-hoc | Advanced | Spatial | Calculations | Statistics

6. **Collaboration**
   - Alerting | Subscriptions | Storytelling | Sharing | Discussions

Deployment Options:
- ON-PREMISES | CLOUD | HOSTED
- WINDOWS | LINUX | MAC
- MULTI-TENANT

Security & Compliance
Extensibility & APIs
Cloud Platform
Big Data is just "DATA"!

Data that swamps traditional approaches of dealing with it
3Vs of DATA Coexist with the 3Vs of Use Cases

Data Volume

Note that all can have different volumes
Principles of a Modern Data Architecture
Key Principles

Support multiple outcomes, applications and personas

Support multiple sources of data

Support multiple tiers of data

Decouple storage from compute

Create an agile data pipeline
Support Multiple Outcomes, Applications and Personas
Personas

Business Users

Data Analysts

Data Scientists
Support Multiple Sources of Data
Types of Data
<table>
<thead>
<tr>
<th></th>
<th>Batch Processing</th>
<th>Stream Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data scope</td>
<td>Queries or processing over all or most of the data</td>
<td>Queries or processing over data on rolling window or most recent data record</td>
</tr>
<tr>
<td>Data size</td>
<td>Large batches of data</td>
<td>Individual records or micro batches of few records</td>
</tr>
<tr>
<td>Performance</td>
<td>Latencies in minutes to hours.</td>
<td>Requires latency in the order of seconds or milliseconds.</td>
</tr>
<tr>
<td>Analytics</td>
<td>Complex analytics.</td>
<td>Simple response functions, aggregates, and rolling metrics.</td>
</tr>
</tbody>
</table>
Support Multiple Tiers of Data
Data Lakes
- Store Everything and Anything
- Unknown Questions with Unknown Answers
- Unstructured/Data Mining/Data Science

Data Warehouses
- Analytical Queries
- Known Questions with Unknown Answers
- Regularly Refreshed Business Concepts

Analytical Databases
- Precomputed Aggregates
- Known Questions and Known Answers
- Analytical Dashboard Already Constructed

Large data (raw or prepared) | Prepared data | Aggregated data | Performance
Velocity of Decision Making

FARM TO TABLE
Decouple Storage from Compute
Decoupling Compute from Storage

The BigQuery Storage Hierarchy

User-level Features
- Public Datasets
- Commercial Datasets
- Marketing Datasets
- Ads DataHub
- IAM and Data Sharing
- Logical Storage Layer
- Capacitor
- Automated DBA Tasks
- Dynamic Execution
- Colossus
- Google Infrastucture

Managed by Google

Unlike on-prem or EC2-fleet Hadoop, EMR Decouples Storage & Compute

Traditional Hadoop
- Tightly-coupled compute & storage
- inflexibility

Amazon EMR
- Decoupled compute & storage
- flexible storage
- Right-sizing compute
Customer Examples
Hadoop (HDFS Storage)

CDH Data lake

Production DBs
Click stream logs
3rd party APIs
Flat files

atScale

Warehouse

Tableau Server + Desktop

SQL + Feeds + CRM + Other

Other
Redshift
Criteria
- Fast analytic queries
- OnDemand reports
- SQL
- Client and Reporting tool support
- Scalable
Prototype

Diagram:
- On-Premises SQL Server
- Source Files (JSON)
- AWS DynamoDB ETL Workflow
- Amazon Redshift
- Landing Zone
- Data Lake
- ODS
- Stage Redshift
- BIF/MFEE AWS Data Pipeline
Updated messaging platform with Amazon Kinesis

Added Amazon Athena to our analytics ecosystem in support of discovery and ad-analyses

New configuration helps us maximize benefit while reducing costs without jeopardizing speed or performance

Successfully deployed Tableau Server to ~1,900 internal users in effort to scale self-service analytics
High Level Reference Architecture
Vendor Big Data Stacks

Data Lake Architecture

[Diagram showing the process of data handling through sources, raw data, compute for prep, storage, compute for query, and analytics, with Tableau as the analytics tool.]

Your Big Data Analytics Platform
Create an Agile Data Pipeline
Create an Agile Data Pipeline
Demo
Sources and Acknowledgements
Sources and Acknowledgements


NYC Taxi - Shape Files: https://github.com/toddwschneider/nyc-taxi-data

Python Modules:

GIS - Fiona / Shapely / Rtree

Redshift - Psycopg2

Blog Posts:

NYC Taxi Data – Inspiration: Todd Schneider


AWS EMR with Presto CLI Scripts: Mark Litwintschik

http://tech.marksblogg.com/billion-nyc-taxi-rides-presto-emr.html

Documentation:

AWS (EMR, Redshift, Spectrum Etc.)

Tableau
Recap Takeaways
General

Embrace continuous improvement
Promote Ad Hoc Analyses & Automate what you can
Design and Redesign as needs evolve

Consider the breadth and depth of use cases
Data is multi-channel, multi-layer, even multi sourced
Don’t “always” wait for it to be in the lake.

Learn from the community
You will hear about successes this week.

One size doesn't fit all…
Create Stored Data for Analysis:
Please complete the session survey from the Session Details screen in your TC18 app.
Thank you!

Contact or CTA info goes here
TABLEAU
CONFERENCE