

Pivotal®

Modern Software Approaches for Operationalizing the Application of Machine Learning/AI



November 2018

Agenda

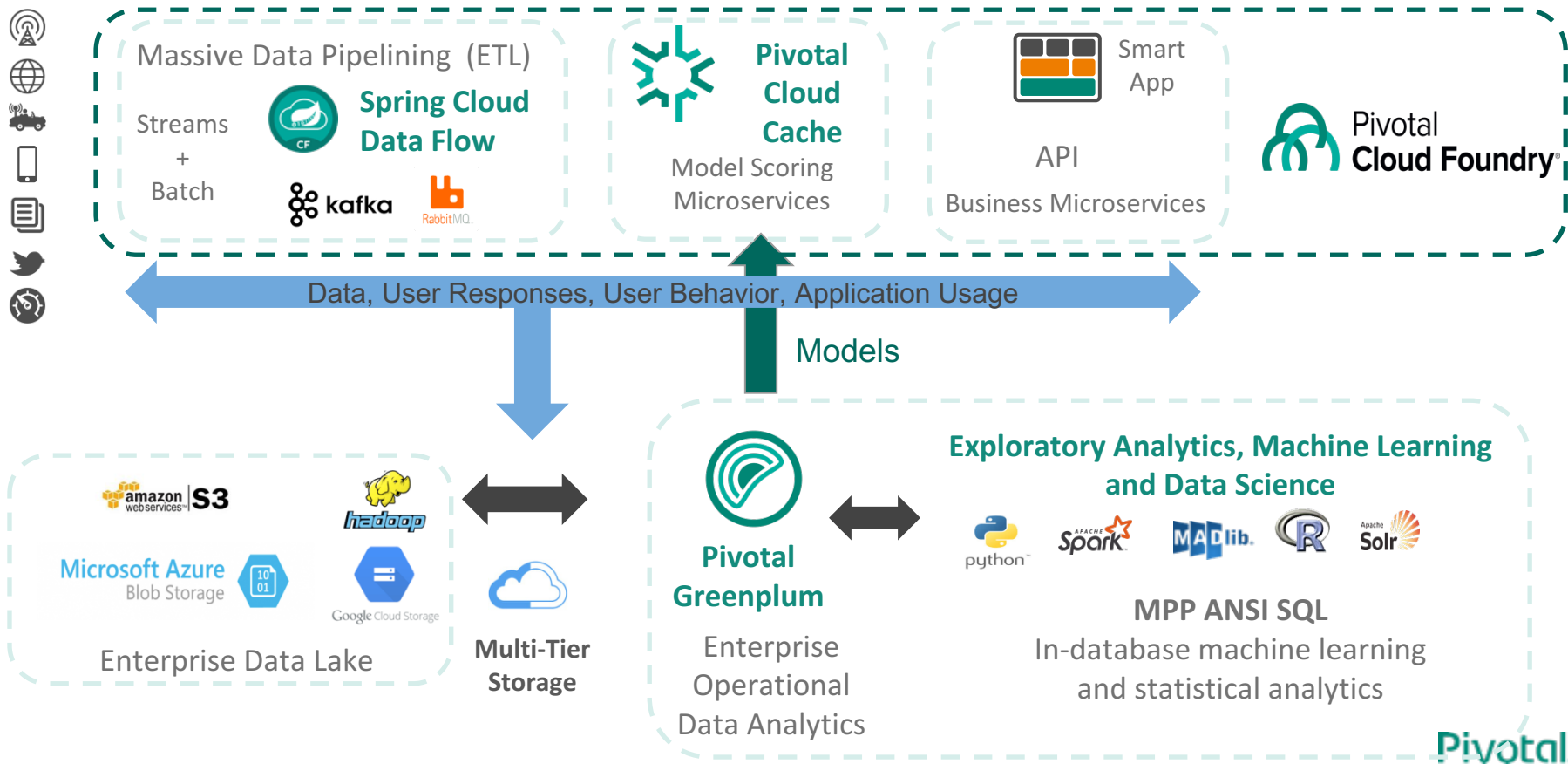
- Introductions
- Process
- Platform
- Analytics
- Discussion

Transforming the Way the World Builds Software



Process - Platform - Analytics

Integrated analytics and Smart Apps



Process

How we build the right software quickly and efficiently

Extreme Programming (XP)

- Key Practices
 - Balanced Team
 - Stories
 - Test Driven Development
 - Pair Programming
 - Automate Everything
 - Sustainable Pace

fedcoop

RRFORCE
DEFENSE

TECH

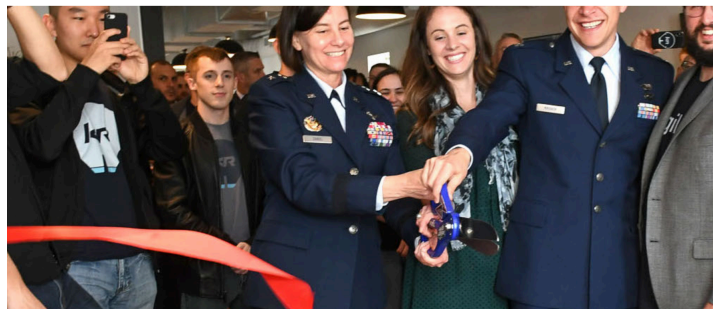
DEFENSE

ACQUISITION

WATCH

LISTEN

Air Force looks to rapidly develop software with Project Kessel Run



Balanced Team

- Product Manger
- Customer
- Designer
- Polyglot Developers

Stories

- Decompose the product into small pieces of functionality
- Typically, implementing these stories is a 1-3 day task.
- Reduces the risk of implementation
- Use technical spikes to determine proper technical direction
- Pivotal Tracker to document and track the stories
 - Tracks velocity to estimate when product will be completed

Test Driven Development

- Tests are written before code
- Tests must fail before implementation begins
- Implementation is complete when all tests succeed

Pair Programming

- Two monitors, 1 keyboard
- Continuous code review
- Teamwide understanding of how the code works
- Learning by collaboratine

Automate Everything

- Continuous Integration
- Fully Automated Deployment to Production
 - Circuit breakers
 - Blue-green deployments

Sustainable Pace

- Work 9 AM to 6 PM with 1 hour for lunch
 - Consistent team collaboration
- Manage to deliver "on time"
 - Definition of MVP
 - Pair programming allows new members at any time

Extreme Programming (XP)

- What our customers say
 - 40% reduction in time to delivery
 - Products that meet customer needs
 - Continuous evaluation of the product features
 - “Customer” input daily



07.05.18

The U.S. Air Force learned to code—and saved the Pentagon millions

In partnership with Pivotal Labs, a pilot program is out to remake how the Pentagon acquires weapons systems.



Platform

How can we efficiently and reliably deploy and manage applications?

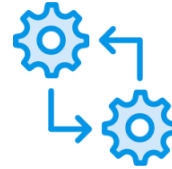
**Here is my source code
Run it in the cloud for me
I do not care how**

- Onsi Fakhouri

One size does not fit all...



CONTAINERS



Batches



EVENT-DRIVEN
FUNCTIONS



MICROSERVICES



DATA
SERVICES/COTS



MONOLITHIC
APPLICATIONS

Container
Orchestrator
(CaaS)

Application
Platform
(PaaS)

Serverless
Functions
(FaaS)

IaaS

Pivotal Application Service (PAS): A Runtime for Apps



Increase speed and deploy code to production thousands of times per month. Use PAS to run Java, .NET, and Node apps.

Best runtime for Spring and Spring Boot — Spring's microservice patterns—and Spring Boot's executable jars—are ready-made for PAS.

Turnkey microservices operations and security — Spring Cloud Services brings microservices best practices to PAS. It includes Config Server, Service Registry, and Circuit Breaker Dashboard.

A native Windows and .NET experience — Use PAS to run new apps built with .NET Core. Run your legacy .NET Framework apps on PAS too, using the .NET Hosted Web Core buildpack. Push applications to containers running on Windows Server 2016.

Built for apps — PAS has everything to need to run apps. Buildpacks manage runtime dependencies; metrics, logging, and scaling are done for you. Multitenancy, and blue/green deployment patterns are built-in. Extend apps with a rich service catalog.

Container-ready — PAS supports the OCI format for Docker images. Run platform-built and developer-built containers.

Our Success Measures - the 5 Ss Sustained

Speed

Operators can efficiently manage the platform and onboard new teams.

Developers can iterate on delivering consumer value rapidly.

Stability

Operators can reason about the stability of the platform and provide well-understood SLOs.

Developers can rely on the platform to allow them to deliver outcomes with low volatility.

Scalability

Operators can provide a platform that meets their scale needs.

Developers can ramp productivity linearly with personnel.

Developers can run applications that handle large-scale loads.

Security

Operators can trust a secure-by-default platform that solves their security needs without introducing toil.

Developers experience the safety to experiment and iterate rapidly..

Savings

Operators can serve thousands of devs within tight budget constraints.

Operators have choice around which cloud to run on.

Developers reduce waste through small batch delivery and fast feedback.

Sustained

The platform can deliver on all of these outcomes as efficiently on day 1000 as it does on day 1.

Examples

- Scale application instances
 - Command line
 - Fully automated
- Zero downtime upgrades
 - Use blue-green deployment



20,000 containers

4 ops people

50% reduction in mean time to repair

88% reduction in downtime

10 deploys/yr --> **130** deploys/yr

iPhone X Launch:

T-Mobile



- No limitations,
- No shopping experience mishaps,
- No crashes!
- T-Mobile Uses **Cloud Foundry** and **Cloud Cache**

All other providers had significant site outages or long delays.

Analytics

Operationalizing Training and Inference

Massively Parallel Analytics with Greenplum

Unified data in a single environment makes distributed, in-database analytics practical



Unified Data for Expressive Power

Analyze more types of data in a single environment - text, geospatial, graph

In-database parallel supervised and unsupervised methods

Petabyte scale for long-tail insights



In-Database Analytics for Speed

Distributed ML reveals hidden anomalies faster

MPP architecture trains more models in less time

Facilitates ensemble modeling for more predictive power



Adapt Quickly to Changing Data

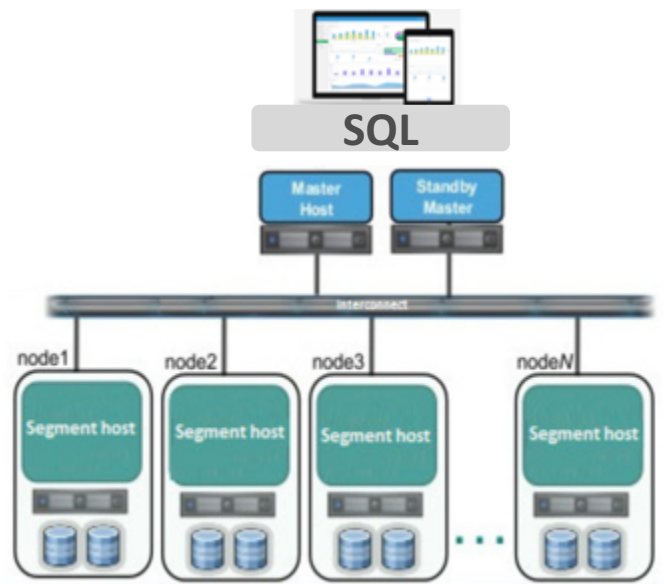
Environments and data patterns change, making models obsolete

MPP enables rapid modeling of individuals and entities at scale

Familiar SQL + MPP accelerates common data prep, quality tasks

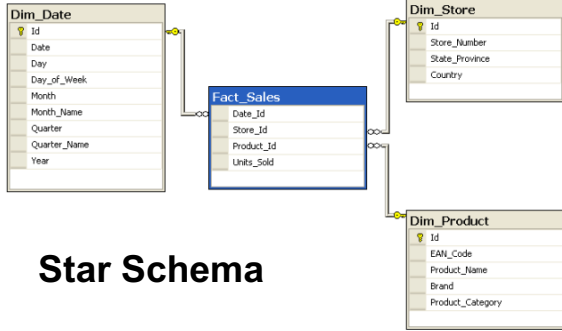
Greenplum Data Platform

Flexible framework for processing large datasets

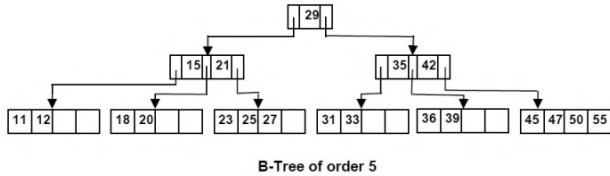


- **Web-scale architecture:** Highly distributed, shared nothing software
- **Extreme performance:** parallel, linear scale-out
- **Unified data platform**
 - structured, unstructured data, external data access
 - polymorphic storage support
- **Simple and automatic**
 - Just load and query like any database
 - Tables are automatically distributed across nodes
- **Petascalse Loading** - Parallel Load & Unload
- **Infrastructure Agnostic**
- **Optimized for Business Intelligence and Analytics**

Multiple Data Formats & Storage Locations Support



Star Schema



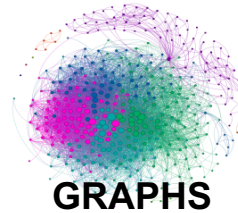
B-Tree Indices



Apache ORC™



JavaScript Object Notation



Greenplum Hadoop & Cloud Connectors



Operational
Analytics & SQL



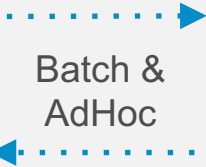
SLA Driven &
Iterative




Pivotal
Greenplum




Data Lake &
Cold Storage




Batch &
AdHoc




Hortonworks




S3




cloudera



Google Cloud Platform



MAPR



Azure

Public & Private
Data Lakes

Warm

Data Temperature

Cold

ML, Statistical, Graph, Path Analytics

Generalized Linear Models

- Linear Regression
- Logistic Regression
- Multinomial Logistic Regression
- Ordinal Regression
- Cox Proportional Hazards Regression
- Elastic Net Regularization
- Robust Variance (Huber-White), Clustered Variance, Marginal Effects

Utility Modules

- Array and Matrix Operations
- Sparse Vectors
- Random Sampling
- Probability Functions
- Data Preparation
- PMML Export
- Conjugate Gradient
- Stemming
- Sessionization
- Pivot
- Path Functions
- Encoding Categorical Variables

Other Machine Learning Algorithms

- Principal Component Analysis (PCA)
- Association Rules (Apriori)
- Topic Modeling (Parallel LDA)
- Decision Trees
- Random Forest
- Conditional Random Field (CRF)
- Clustering (K-means)
- Cross Validation
- Naïve Bayes
- Support Vector Machines (SVM)
- Prediction Metrics
- K-Nearest Neighbors

Matrix Factorization

- Singular Value Decomposition (SVD)
- Low Rank

Linear Systems

- Sparse and Dense Solvers
- Linear Algebra

Descriptive Statistics

Sketch-Based Estimators

- CountMin (Cormode-Muth.)
- FM (Flajolet-Martin)
- MFV (Most Frequent Values)

Correlation and Covariance

Summary

Inferential Statistics

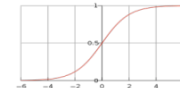
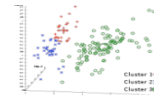
- Hypothesis Tests

Time Series

- ARIMA

Graph

- PageRank
- Single Source Shortest Path



Graph Analytics

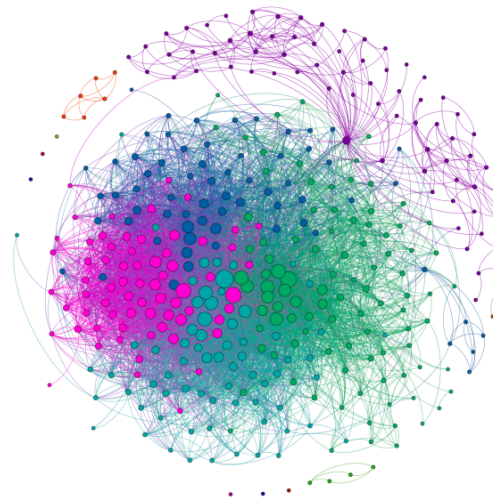
Designed for very large graphs
(*billions* of vertices/edges)

No need to move data and
transform for external graph engine

Familiar SQL interface

Algorithms:

- All pairs shortest path*
- Breadth first traversal*
- Connected components*
- Multiple graph measures*
- PageRank
- Single source shortest path



```
SELECT madlib.pagerank(  
  'vertex',          -- Vertex table  
  'id',             -- Vertex id column  
  'edge',           -- Edge table  
  'src=src, dest=dest', -- Comma delimited string of edge arguments  
  'pagerank_out',  -- Output table of PageRank  
  NULL,            -- Default damping factor (0.85)  
  NULL,            -- Default max iters (100)  
  0.00000001,     -- Threshold  
  'user_id');     -- Grouping column name
```

Vertex Table

Vertex	Vertex Params	...
0	...	
1	...	
2	...	
3	...	
.		
.		

Edge Table

Source Vertex	Dest Vertex	Edge Weight	Edge Params	...
0	3	1.0	...	
1	0	5.0	...	
1	2	3.0	...	
2	3	8.0	...	
3	0	3.0	...	
3	1	2.0	...	
.				
.				

Text Analytics



GPText: SQL Warehousing + Text

- Leveraging Apache Solr and GPDB
- 5 years commercial production experience
- Madlib integration for machine learning on text data
- PL/Python and PL/Java integration for Natural Language Processing

Use Cases

- Communications compliance and monitoring
- Customer Sentiment analysis
- Document Search and Query
- Social Media Processing, etc.



GeoSpatial Analytics



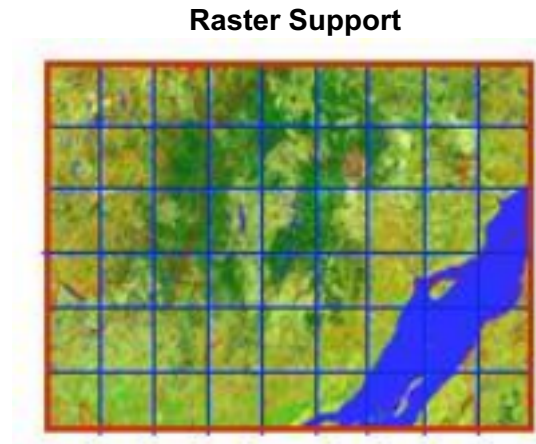
Current Key Features:

- Points, Lines, Polygons, Perimeter, Area, Intersection, Contains, Distance, Long/Lat

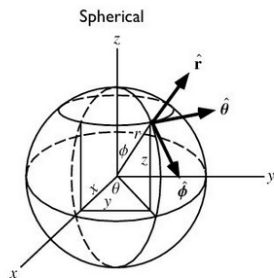
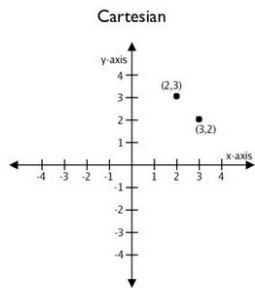
```
geodemo=# SELECT
nyc_subway_stations.long name AS subway,
nyc_neighborhoods.name AS neighborhood
FROM nyc_neighborhoods
JOIN nyc_subway_stations
ON ST_Contains(nyc_neighborhoods.geom, nyc_subway_stations.geom)
WHERE nyc_neighborhoods.name = 'Greenwich Village';
```

subway	neighborhood
W 4th St (B,D,F,V) Manhattan	Greenwich Village
14th St / Union Sq (4,5,6) Manhattan	Greenwich Village
14th St (1,2,3) Manhattan	Greenwich Village
Bleecker St / Broadway-Lafayette St (6) Manhattan	Greenwich Village
Christopher St / Sheridan Sq (1) Manhattan	Greenwich Village
Union Sq / 14th St (L,N,O,R,W) Manhattan	Greenwich Village
6th Ave / 14th St (F,L,V) Manhattan	Greenwich Village
8th St / New York University (N,R,W) Manhattan	Greenwich Village
Astor Pl (6) Manhattan	Greenwich Village
W 4th St (A,C,E) Manhattan	Greenwich Village

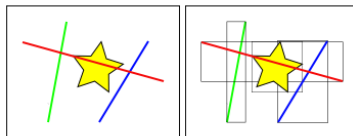
(10 rows)



Round earth calculations



Spatial Indexes & Bounding Boxes



Products

Work with Boundless to Create Your Next GIS Application

Social Media Globalization

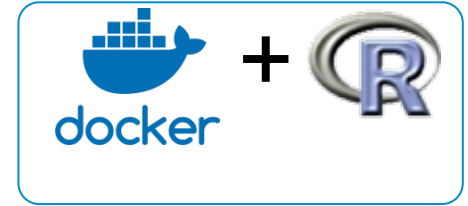
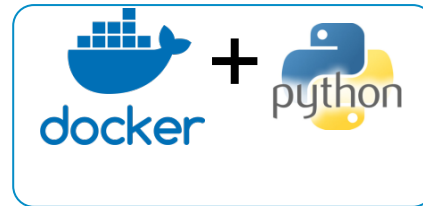
Boundless offers support, maintenance, training and professional services.

[Learn More](#)

Procedural Language Extensions (PL/X)



- Current Computing Interfaces
 - User Defined Types
 - User Defined Functions
 - User Defined Aggregates
- PL/Container - isolate and customize processing in containers for PL/languages (in beta now)

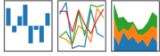


Native Support for R/Python Libraries (Partial List)



pandas

$$y_i = \beta^T x_i + \mu_i + \epsilon_i$$



gensim



NumPy

MCMCpack



TensorFlow

pyLDAvis



spaCy

LIFELINES



XGBoost

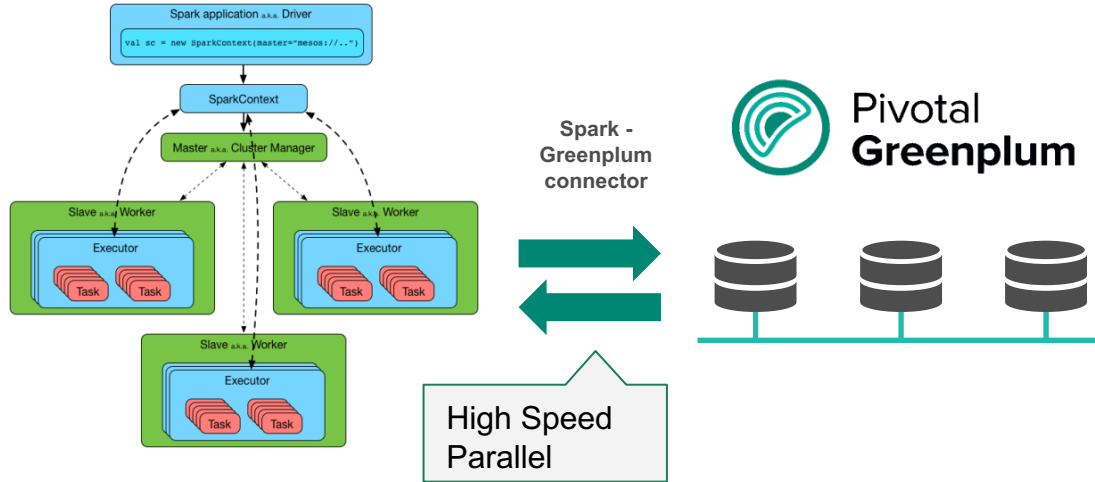
BeautifulSoup



Greenplum-Spark Connector



In-memory processing



- Provide Data Access to Greenplum Data
- Leverage SPARK Skill Set of Data Scientists
- Leverage off-cluster compute resources to do computations
- Push result sets back into Greenplum for storage

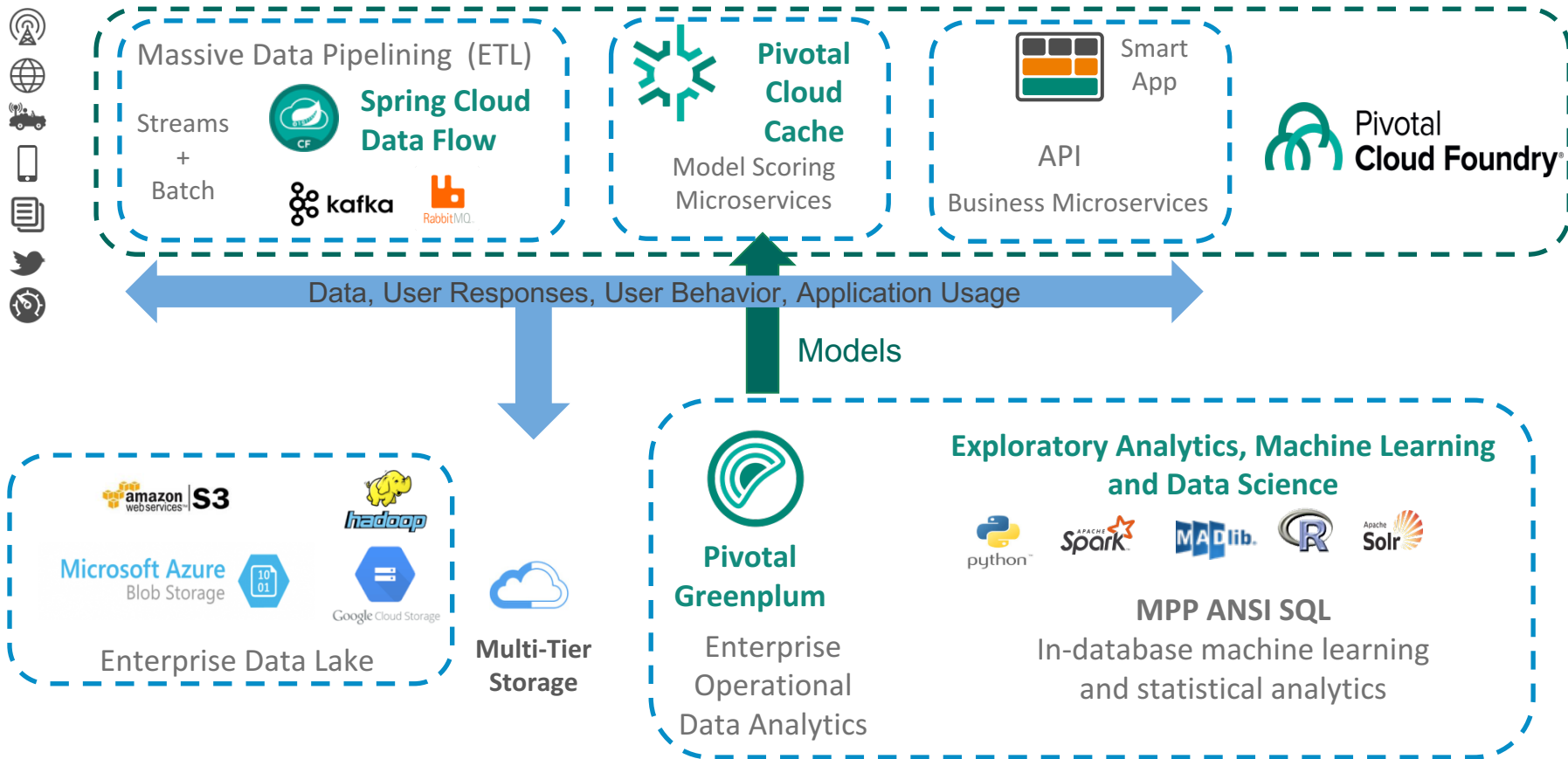
Examples

- Churn Prevention
 - Fully automated data governance delivers data to team
 - Over 200 models in production automated training and scoring
- Fraud Detection
 - First model in production in 6 weeks (version 0.1)
 - Virtually eliminated a specific type of fraud

Closed Loop Analytics

Enabling agile delivery of analytics driven applications

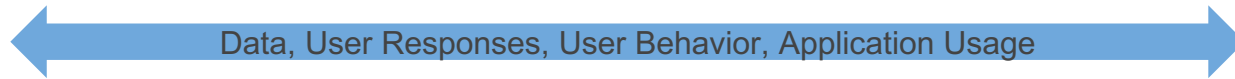
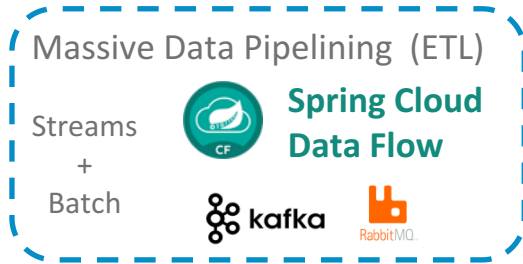
Integrated analytics and Smart Apps



Integrated analytics and Smart Apps



Integrated analytics and Smart Apps



Integrated analytics and Smart Apps



Massive Data Pipelining (ETL)

Streams
+
Batch



Spring Cloud
Data Flow



kafka



RabbitMQ

Data, User Responses, User Behavior, Application Usage

amazon S3
web services



hadoop

Microsoft Azure
Blob Storage



Google Cloud Storage



Multi-Tier
Storage

Enterprise Data Lake



Pivotal
Greenplum

Enterprise
Operational
Data Analytics

Exploratory Analytics, Machine Learning
and Data Science



python



APACHE
Spark



MADlib



R



Apache
Solr

MPP ANSI SQL

In-database machine learning
and statistical analytics

Integrated analytics and Smart Apps



Massive Data Pipelining (ETL)

Streams
+
Batch



Spring Cloud
Data Flow



kafka



RabbitMQ



Pivotal
Cloud
Cache

Model Scoring
Microservices

Data, User Responses, User Behavior, Application Usage

amazon S3
web services



hadoop

Microsoft Azure
Blob Storage



Google Cloud Storage



Multi-Tier
Storage

Enterprise Data Lake



Pivotal
Greenplum

Enterprise
Operational
Data Analytics

Exploratory Analytics, Machine Learning
and Data Science



python



APACHE
Spark



MADlib



R



Apache
Solr

MPP ANSI SQL


In-database machine learning
and statistical analytics



Integrated analytics and Smart Apps




Massive Data Pipelining (ETL)


Streams
+
Batch

 **Spring Cloud Data Flow**

 **kafka**  **RabbitMQ**

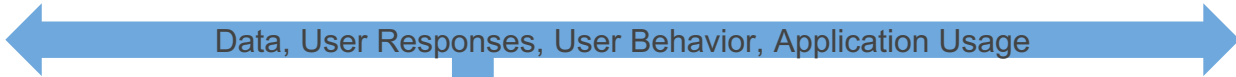
 **Pivotal Cloud Cache**

Model Scoring
Microservices

 **Smart App**

API

Business Microservices




 **amazon S3**  **hadoop**

 **Microsoft Azure Blob Storage**  **Google Cloud Storage**






Enterprise Data Lake

 **Multi-Tier Storage**

 **Pivotal Greenplum**

Enterprise Operational Data Analytics

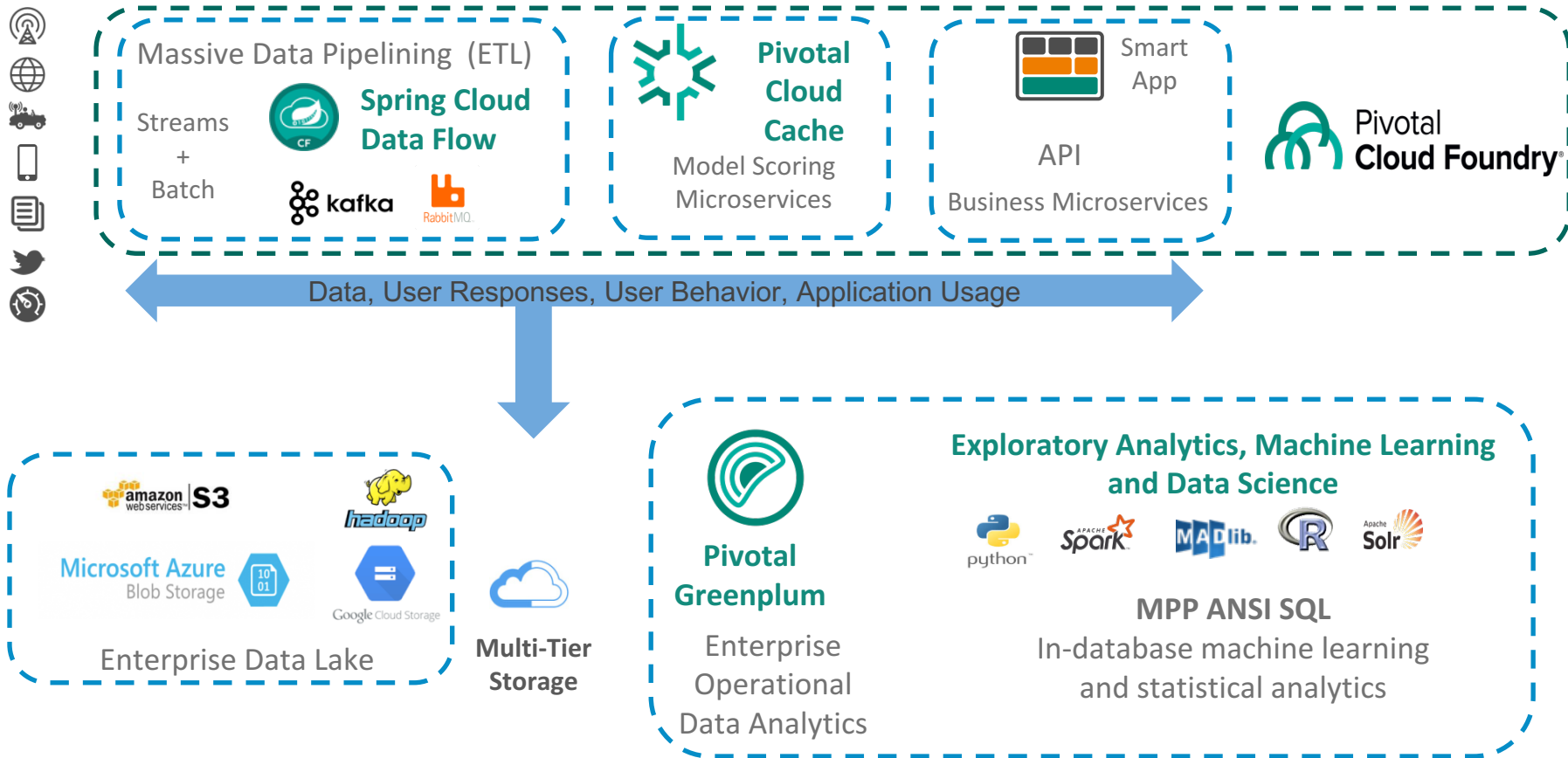
Exploratory Analytics, Machine Learning and Data Science

 **python**  **APACHE Spark**  **MADlib**  **R**  **Apache Solr**

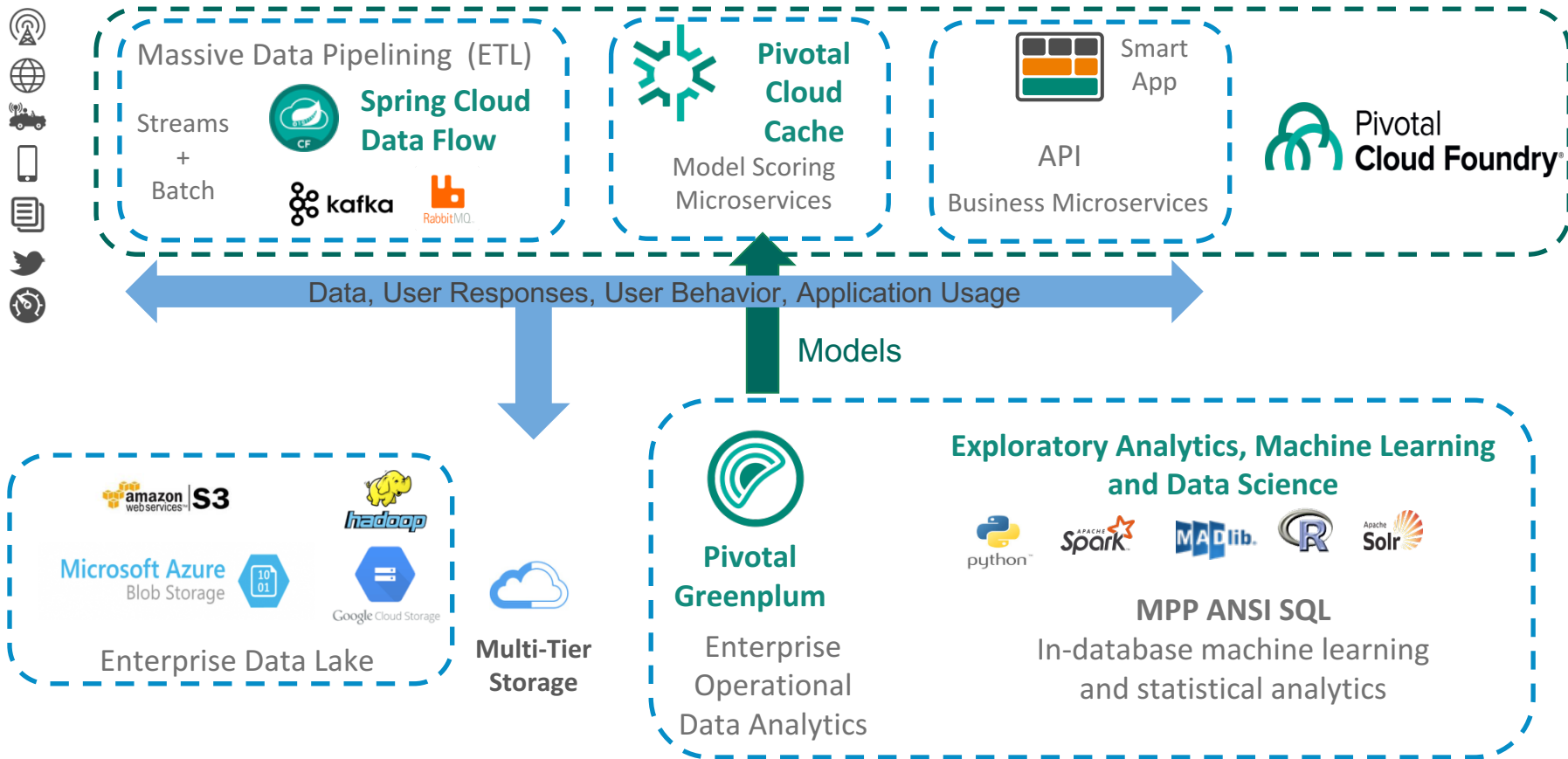
MPP ANSI SQL

In-database machine learning and statistical analytics

Integrated analytics and Smart Apps



Integrated analytics and Smart Apps



Discussion

Transforming the Way the World Builds Software



Process - Platform - Analytics