Data Science @Pivotal

Kaushik K. Das
Senior Principal Data Scientist
How do you make an oil platform smart enough to prevent a disaster?

#1 – Macondo experience focuses industry on prevention and incident management

Drivers
- Investigations continue to reveal failure points
  - Critical decisions on delay of blow out equipment maintenance
  - Speed of capping off, with multiple functions performed at once
  - Failure of safety protocols and multiple alarming systems
- Review of containment and spill response

Predictions
- The industry will respond to the incident by developing safety protocols and training, triage of alarms, and automated shut-downs. Decision-making and lines of authority will be better defined
- New regulatory requirements will drive increased investment by oil and gas companies in Health, Safety and Environment (HSE) system enhancements, upgrades, and replacements. Incident tracking and management, corrective action and regulatory reporting will be addressed first and will depend on application integration
- Advanced analytics such as neural networks, self-organizing maps, data mining, and predictive analytics will be used to predict potential drilling failures and simulate post-incident scenarios. 4-D visualization will be shared between well planners and drillers.
Smart System = Physical Machine + Digital Brain
Predictive Analytics with Real-time Ingest

High Frequency & High Concurrency Drill Data

Threshold detection with push notification

Historical Seismic and Geological Data

Real-time Ingest

Smart Drill Platform

f(x)

GEMFIRE

SQLFire

GREENPLUM
Smart Drilling Platform: Workflow

**Objective:** Predict machine parameters based on incoming variable

**Business Impact:** Increased productivity and reduction of time wasted. High accuracy of drill parameter prediction

**Data:**
- Historical well and reservoir data
- Well log data
- Real-time drill data

**Data Gathering:** Collect and store data from sensors, equipment, and reservoirs

**Data Ingestion:** Low latency collection and storage of all operational data from drilling

**Drill Rate Penetration Optimization:** Scoring algorithm on incoming data and anomaly detection against defined model

- Apps power businesses, and those apps generate data
- Analytic insights from that data drive new app functionality, which in-turn drives new data
- The faster you can move around that cycle, the faster you learn, innovate & pull away from the competition
Pivotal’s Opportunity

✔️ Uniquely positioned to help enterprises modernize each facet of this cycle today

✔️ Comprehensive portfolio of products spanning Apps, Data & Analytics

✔️ Converging these technologies into a coherent, next-gen Enterprise PaaS platform
Anatomy of Transformation
IT TAKES MORE THAN ONE TOOL
Evolving Enterprise Data ARCHITECTURE

Data Ingestion System
- Stream/CEP

Data Staging Platform
- Hadoop

Analytic Data Marts
- MPP Database

Operational Intelligence
- In-Memory DB

Run-Time Applications
- In-Memory Object

Enterprise Data Warehouse
- RDBMS

Traditional BI/Reporting
- Data Visualization
PIVOTAL Product Portfolio Today

Data Ingestion System

Data Staging Platform

Analytic Data Marts

Operational Intelligence

Run-Time Applications

Enterprise Data Warehouse

Traditional BI/Reporting Data Visualization

© Copyright 2013 Pivotal. All rights reserved.
In-Database Analytics: Detail
MADlib In-Database Analytics Functions

Data Modeling

- Supervised Learning
  - Naive Bayes Classification
  - Linear Regression
  - Logistic Regression
  - Multinomial Logistic Regression
  - Decision Tree & Random Forests
  - Support Vector Machines
  - Cox Proportional Hazards Regression
  - Conditional Random Field

- Unsupervised Learning
  - Association Rules
  - k-Means Clustering
  - SVD Matrix Factorization
  - Low ranked Matrix Factorization
  - Parallel Latent Dirichlet Allocation

Descriptive Statistics

- Sketch-based Estimators
  - CountMin (Cormode-Muthukrishnan)
  - FM (Flajolet-Martin)
  - MFV (Most Frequent Values)

Support Modules

- Profile
- Quantile

Array Operations
- Conjugate Gradient
- Sparse Vectors
- Random Sampling
- Probability Functions
Integrated with Tools/Languages, incl. R

- Load PivotalR Library

- List the columns in the table and preview the first 3 rows of data (the limit is passed through to the db)

- Examine the resulting model

```r
> library(RPostgreSQL)
> library(PivotalR)
> db.connect()
Created a connection to database with ID 1
[1] 1
> db.objects("public.houses")
[1] "public.houses"
> houses = db.data.frame("public.houses")
An R object pointing to public.houses in connection 1 is created!
> names(houses)
[1] "id"  "tax"  "bedroom"  "bath"  "price"  "size"  "lot"
> preview(houses,3)
   id  tax bedroom bath  price  size  lot
1  2 1050     3 6 1350
2  4 870     2 2 90000 1300 17500
3  6 1350     2 1 90500 820 25700
> n1 = mod1ib.lm(bedroom ~ price + size, houses)
> summary(n1)

MADlib Linear Regression Result

Call:  
mod1ib.lm(formula = bedroom ~ price + size, data = houses)

Coefficients: 
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.460e+00  3.437e-01   4.251  0.001129 **
price      -5.200e-06  3.098e-06  -1.681 0.118523
size       1.353e-03  3.912e-04   0.348 0.729159

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ 1

R-squared: 0.566818
Condition Number: 433006.1
```

- Create the “houses” object as a proxy object in R. The data is not loaded into R

- Run a linear regression. This is executed in-database.

- The model is stored in-database, greatly simplifying the development of scoring applications
The Eightfold Path of Data Science – four phases and four differentiating factors

**Phase 1: Problem Formulation**
Make sure you formulate a problem that is relevant to the goals and pain points of the stakeholders.

**Phase 2: Data Step**
Build the right feature set making full use of the volume, variety and velocity of all available data.

**Phase 3: Modeling Step**
This is where you move from answering what, where and when to answering why and what if?

**Phase 4: Application**
Create a framework for integrating the model with decision making processes and taking action using the Internet of Things.

**Technology Selection**
Select the right platform and the right set of tools for solving the problem at hand.

**Iterative Approach**
Perform each phase in an agile manner, team up with domain experts and SMEs, and iterate as required.

**Building a Narrative**
Create a fact-based narrative that clearly communicates insights to stakeholders.

**Creativity**
Take the opportunity to innovate at every phase.
What is a “Data Scientist”?
Smart System = Physical Machine + Digital Brain

• We are able to create a system that collects the data, learns from it and acts

• And a combination of intelligent sensors, big data platform and data science makes that possible