Data Lake In Action: Real-time, Closed Loop Analytics On Hadoop
Pivotal’s Full Approach

It’s More Than Just Hadoop

Pivotal Data Labs

\[ p(X) = \frac{1}{N} \sum_{n=1}^{N} p_n(X) = \frac{1}{N} \sum_{n=1}^{N} \sum_{w=1}^{W} w_n(X) = \sum_{n=1}^{N} \left( \frac{1}{N} \sum_{w=1}^{W} w_n(X) \right) X \]
Why Pivotal Exists

Solve the Big Data Utility Gap

70% of data generated by customers
80% of data being stored

3% being prepared for analysis
0.5% being analyzed
<0.5% being operationalized

First Movers
Google
~7X $290B
~20X $120B

UBER
~30X $4B
~20X $2.9B

Average Enterprises

Smart Enterprises

© Copyright 2014 EMC Corporation. All rights reserved.
Journey To Data Driven Enterprise

**Steps**

- **Archive**
  - Realize cost efficiencies and extend life of existing systems and
  - Data migration

- **Insights**
  - Integrate all existing data to generate business insights
  - Data Analysis

- **Apps**
  - Build Apps to assist/take (automated) actions from the insights generated
  - Data Driven Apps

- **Business Models**
  - Create new revenue streams leveraging new data and new insights
  - Business Transformation

- **Repeatable Framework**
  - Platform for experimenting data driven business models and innovation
  - Experimentation Platform

**Target Technology**

- Data Lake

**Platform as a Service**

**Manager**

**IT Leaders**

**Business Leader**

**CEO**

© Copyright 2014 EMC Corporation. All rights reserved.
Data Driven: Harder Than It Sounds

Real Time
- Analytical
- Transactional

Operationalize
- Ingest
- Interface
- Distill
- Process

Predictive Call Routing, Fraud Prediction, Dynamic Pricing, Re-Marketing, Stream Analytics

Near Real Time
- Analytical
- Transactional

Operationalize
- Ingest
- Interface
- Distill
- Process

Analytic Model Designs, Transaction Analysis, Trend Analysis

Batch
- Analytical
- Transactional

Operationalize
- Ingest
- Interface
- Distill
- Process

ETL, Archive, Trending, Monthly and Weekly Jobs
Data Driven: Impossible In Silos

Data Growth Over 60% Floods These Silos

Finance | Manufacturing | Marketing | IT

© Copyright 2014 EMC Corporation. All rights reserved.
Pivotal Business Data Lake Architecture

Unified Sources
- Real-time ingestion
- Micro batch ingestion
- Batch ingestion

Centralized Management
- System monitoring
- System management

Unified Data Management Tier
- Data mgmt. services
- MDM
- Audit and policy mgmt.

Processing Tier
- In-memory
- MPP database

Distillation Tier
- HDFS storage
  - Unstructured and structured data

Workflow Management

Flexible Actions
- Real-time insights
- Interactive insights
- Batch insights
Pivotal HD Center Of Data Lake

Pivotal HD Enterprise

HDFS

HBase

HBas e

Virtual Extensions

YARN

Resource Management & Workflow

ZooKeeper

Oozie

Pivotal HD

HAWQ – Advanced Database Services

ANSI SQL + Analytics

MADlib Algorithms

Xtension Framework

Catalog Services

Query Optimizer

Dynamic Pipelining

Pivotal GemFire XD – Real-Time Database Services

ANSI SQL + In-Memory

Distributed In-memory Store

Query Transactions

Ingestion Processing

Hadoop Driver – Parallel with Compaction

Pivotal Command Center

Configure, Deploy, Monitor, Manage

Spring

GraphLab, Open MPI

Pig, Hive, Mahout

Map Reduce

Spring XD

Squoop

Flume

Oozie

Virtual Extensions

YARN

Resource Management & Workflow

ZooKeeper

Oozie

Apache

Pivotal

© Copyright 2014 EMC Corporation. All rights reserved.
Pivotal HD Value

- Cost-based Query Optimizer
- ANSI SQL Compliant
- Linear, incremental scalability on COTS hardware
- Deep Analytic OLAP Queries
- Petabyte Data Storage & Management
- Low latency updates and transactions
- Partitioned Events in situ w/ data
- Active-active deployment across WAN
Data Lake In Action
Real-time, Closed Loop Analytics With Pivotal HD

- Adaptive model development based on long term trends
- Real-time model scoring
- Continuous queries

Market Data → Detect Threshold → HAWQ → Historical Data

Monte Carlo Simulation
Recalculate Model

Fast Active Data + Big Historical Data

Pivotal
## The Scenario Yesterday

<table>
<thead>
<tr>
<th>Application Type</th>
<th>Pivotal Component</th>
<th>Pricing Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong> Database</td>
<td>Greenplum DB</td>
<td>Data storage: tiered terabytes</td>
</tr>
<tr>
<td><strong>2</strong> Hadoop Distributed File System</td>
<td>Pivotal HD</td>
<td>Nodes</td>
</tr>
<tr>
<td><strong>3</strong> Parallel Query Engine</td>
<td>HAWQ</td>
<td>Nodes</td>
</tr>
<tr>
<td><strong>4</strong> In-Memory Data Grid</td>
<td>GemFire</td>
<td>CPUs and Add Ons with restrictions</td>
</tr>
<tr>
<td><strong>5</strong> In-Memory Data Grid for Hadoop</td>
<td>GemFire XD</td>
<td>TBD</td>
</tr>
<tr>
<td><strong>6</strong> In-Memory Data Grid with SQL Layer</td>
<td>SQLFire</td>
<td>CPUs and Add Ons with restrictions</td>
</tr>
</tbody>
</table>

* GemFire XD will be included upon GA- Est. Q2- 2014  
Other add-on products: Pivotal Data Dispatch, Alpine Chorus
# The Scenario Today

<table>
<thead>
<tr>
<th>Application Type</th>
<th>Pivotal Component</th>
<th>Pricing Metric:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Database</td>
<td>Greenplum DB</td>
<td>SKU</td>
</tr>
<tr>
<td>2  Hadoop Distributed File System</td>
<td>Pivotal HD</td>
<td>Unit of Measure</td>
</tr>
<tr>
<td>3  Parallel Query Engine</td>
<td>HAWQ</td>
<td>Price</td>
</tr>
<tr>
<td>4  In-Memory Data Grid</td>
<td>GemFire</td>
<td></td>
</tr>
<tr>
<td>5  In-Memory Data Grid for Hadoop</td>
<td>GemFire XD*</td>
<td></td>
</tr>
<tr>
<td>6  In-Memory Data Grid with SQL Layer</td>
<td>SQLFire</td>
<td></td>
</tr>
</tbody>
</table>

* GemFire XD will be included upon GA. Est Q2-2014
World’s Leading Experts
Pivotal Labs – Pivotal Data Labs

On Demand Services
Pivotal Data Dispatch

Pivotal™ Big Data Suite

- GemFire
  - REAL TIME
- Greenplum DB
  - NEAR TIME
- Pivotal HD
  - BATCH
- GemFire XD
  - REAL TIME
- HAWQ
  - NEAR TIME
Customer Centric Model

- Subscription Based
- Software Only
- Core Based
- Customer Incentives
- Flexible Licensing

UNLIMITED PIVOTAL HD INCLUDED
How Does This Work In Practice?

Store Everything
- Obsessively collect data
- Keep it forever
- Put the data in one place

Analyze Anything
- Cleanse, organize, and manage your data lake
- Make the right tools available
- Use the resources wisely to compute, analyze, and understand data

Build the Right Thing
- Use insights to iteratively improve your product