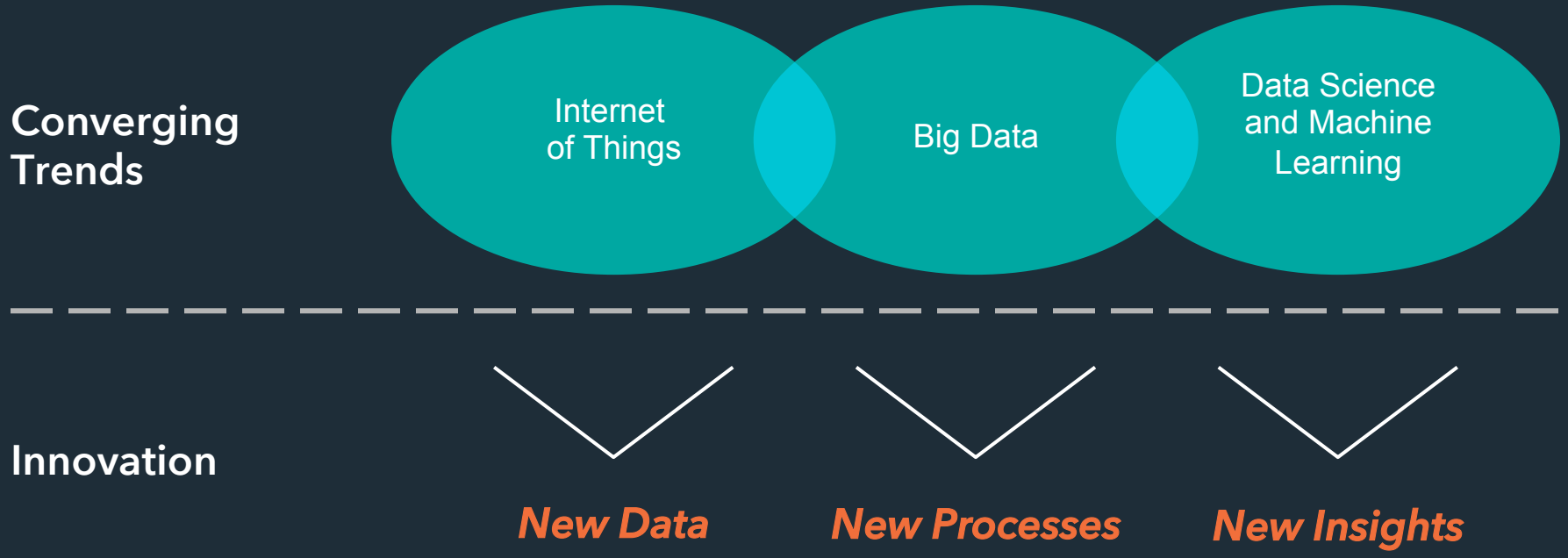
Data Streaming and IoT

Using Pivotal Big Data Suite

Fred Melo
fmelo@pivotal.io
@fredmelo_br

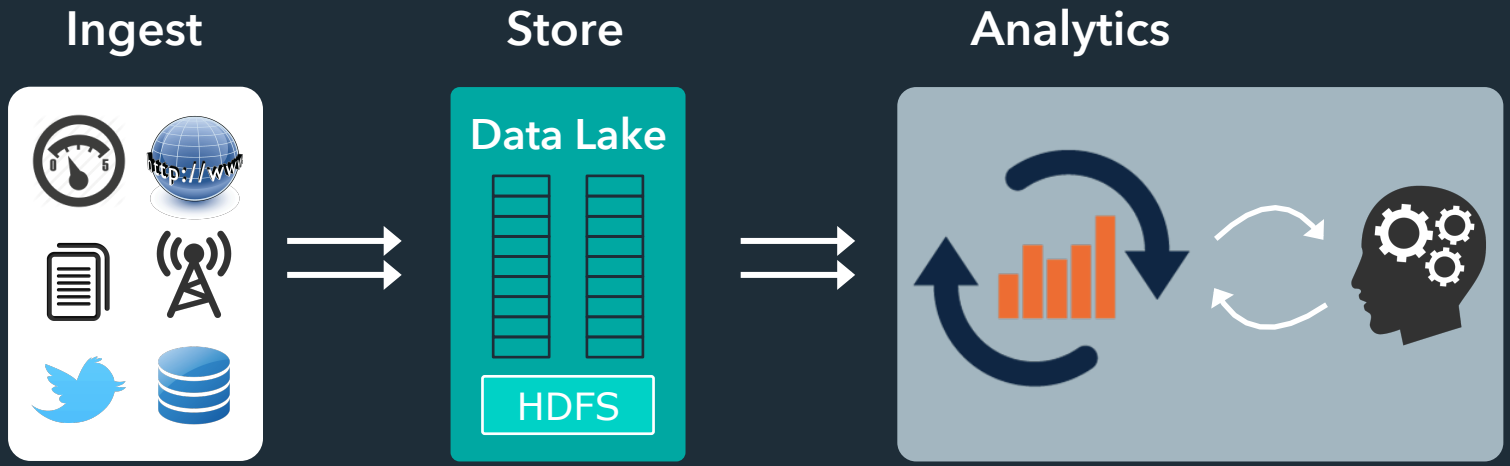
Pivotal

The Journey to the Data-Driven Enterprise



Pivotal

Migrating from a Reactive, Static and Constrained Model…

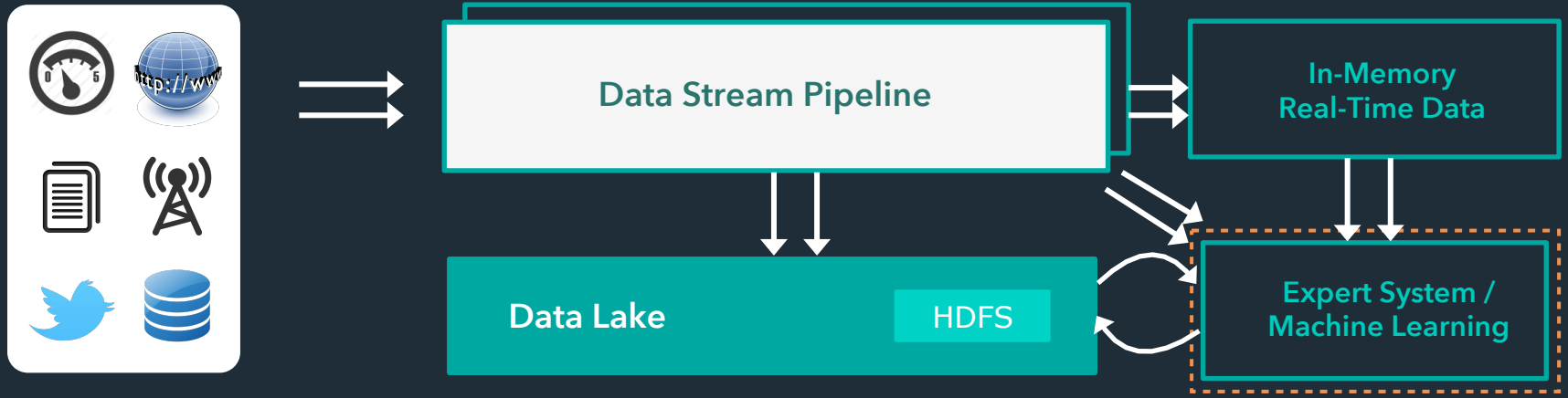


Coding based
No real-time information
Based on expensive ETL

Hard to change
Labor intensive
Inefficient

Pivotal

To Pro-Active, Self-Improving, Machine Learning Systems



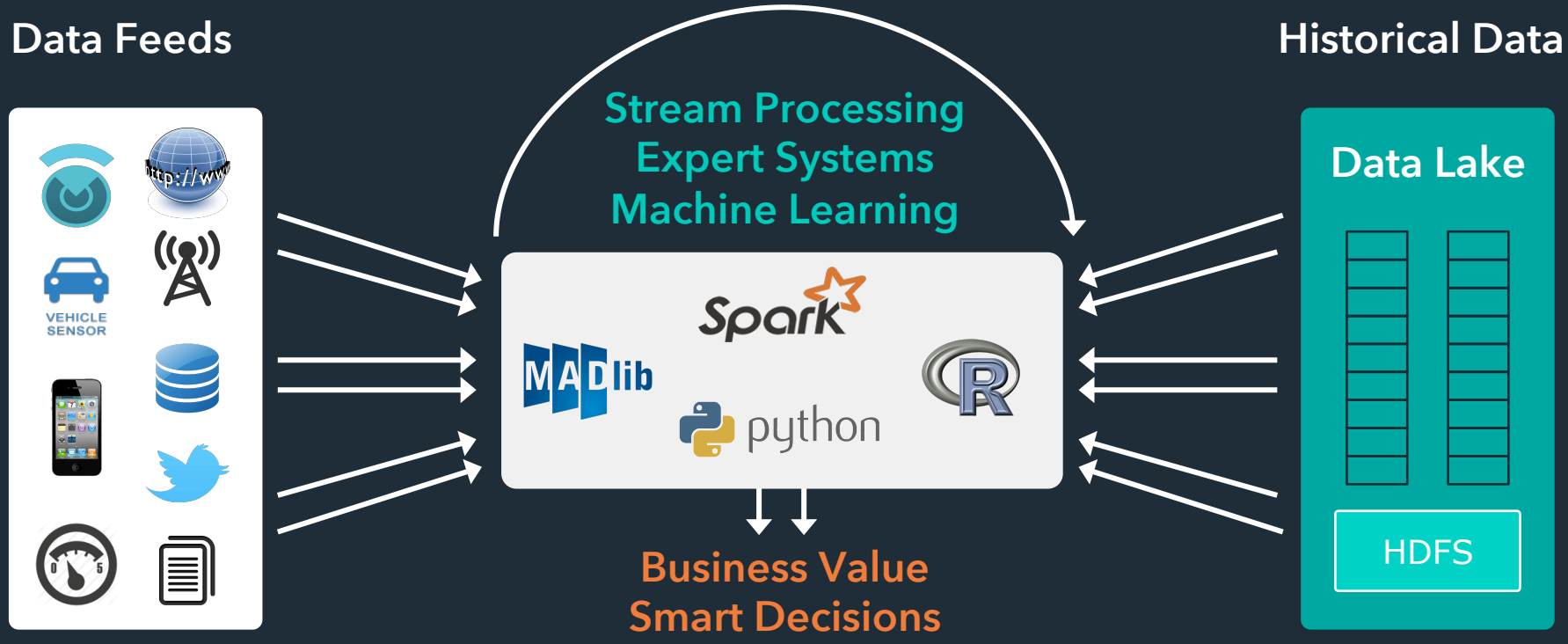
Multiple Data Sources
Real-Time Processing
Store Everything

Continuous Learning
Continuous Improvement
Continuous Adapting

Pivotal

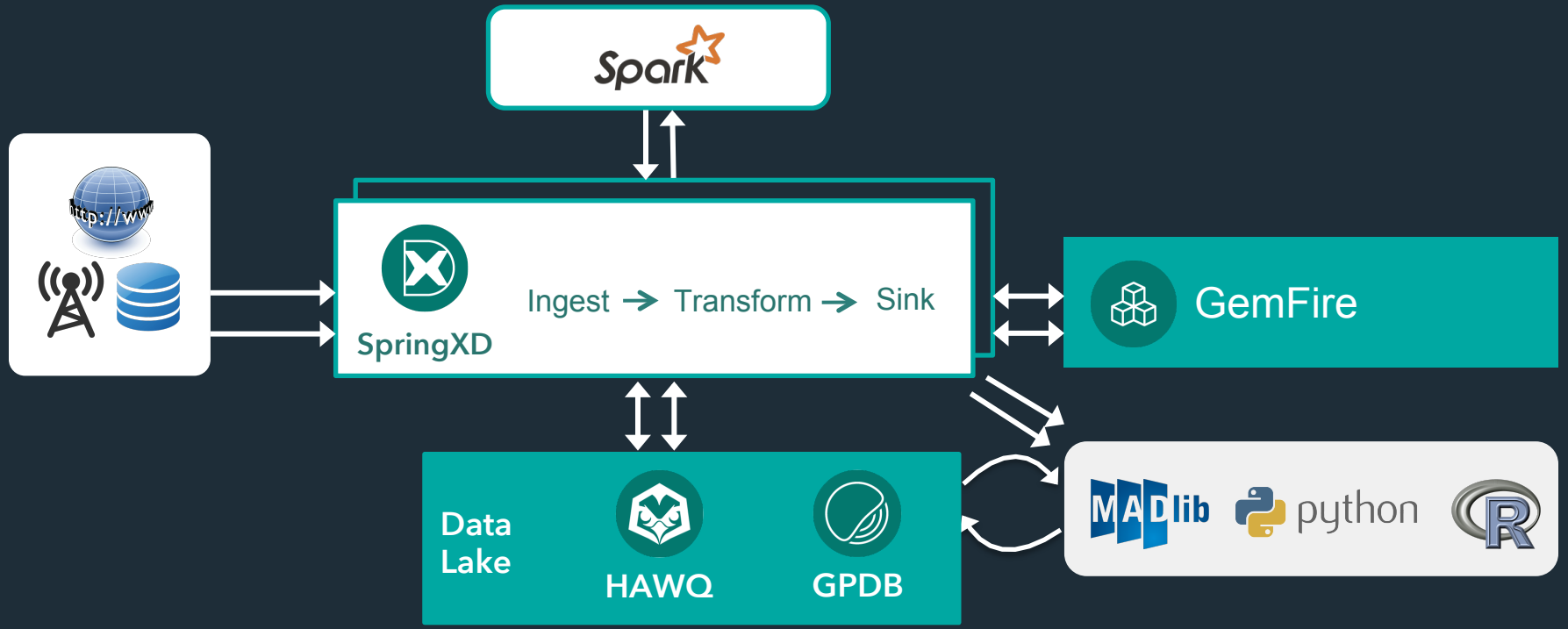
New York Times Research: http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

Still…



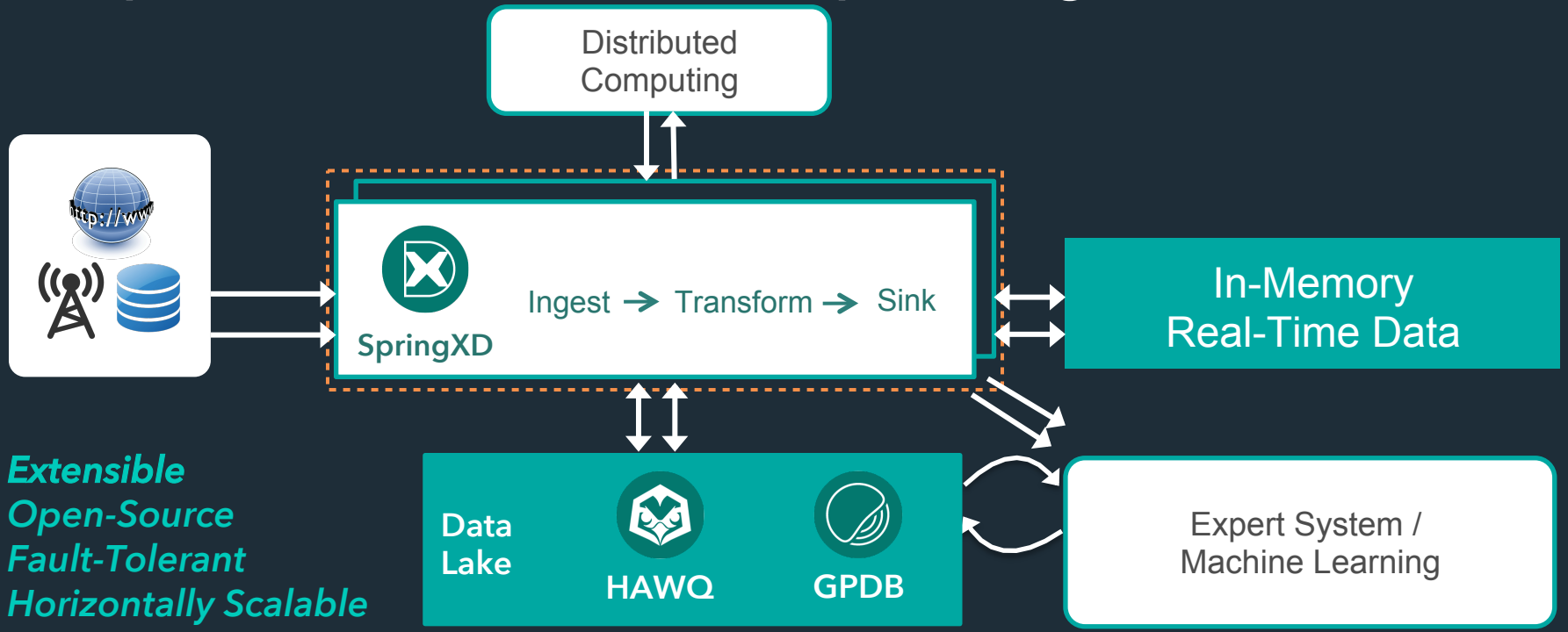
Pivotal

Data Stream Needs an Agile, Scalable and Fast Solution



Pivotal

Spring XD Orchestrates and Automates all the Steps on Data Stream Pipelining



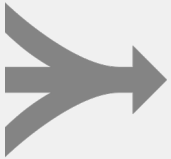
Extensible
Open-Source
Fault-Tolerant
Horizontally Scalable

Pivotal

Spring XD

State of the Art Data Pipeline Automation

INGEST / SINK



- No coding required
- Dozens of built-in connectors
- Seamless integration with Kafka, Sqoop
- Create new connectors easily using Spring

PROCESS

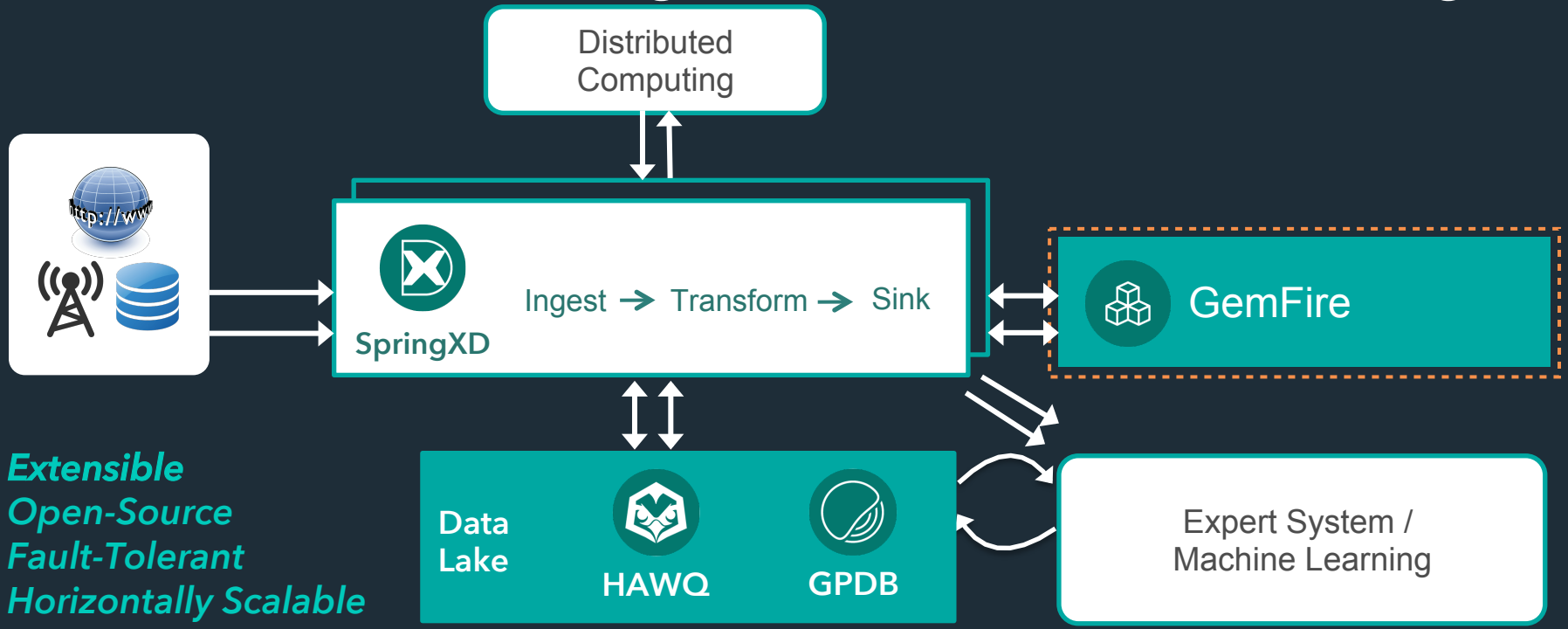
- Call Spark, Reactor or RxJava
- Built-in configurable filtering, splitting and transformation
- Out-of-box configurable jobs for batch processing

ANALYZE

- Import and invoke PMML jobs easily
- Call Python, R, Madlib and other tools
- Built-in configurable counters and gauges

Pivotal

GemFire Provides Scalable, Low-Latency Data Access, Storage and Event Processing



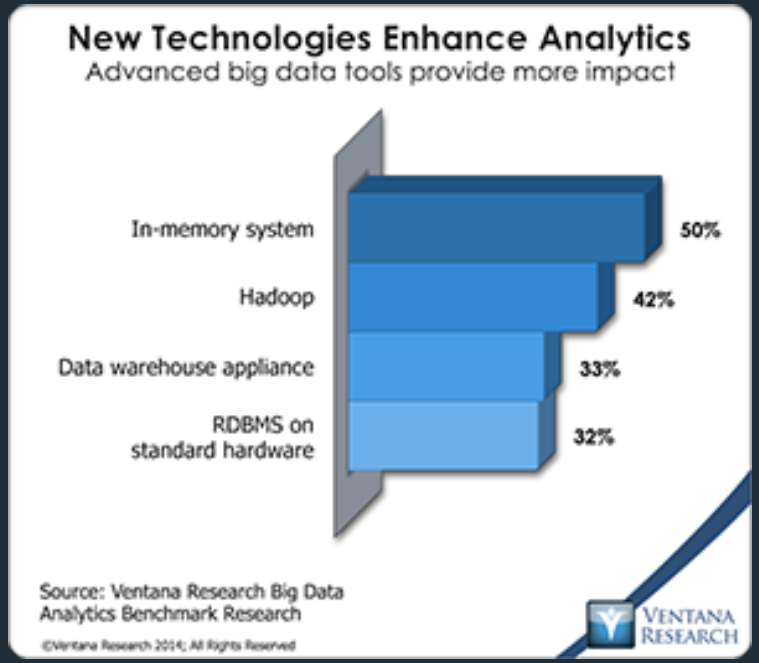
Extensible
Open-Source
Fault-Tolerant
Horizontally Scalable

Pivotal

GemFire

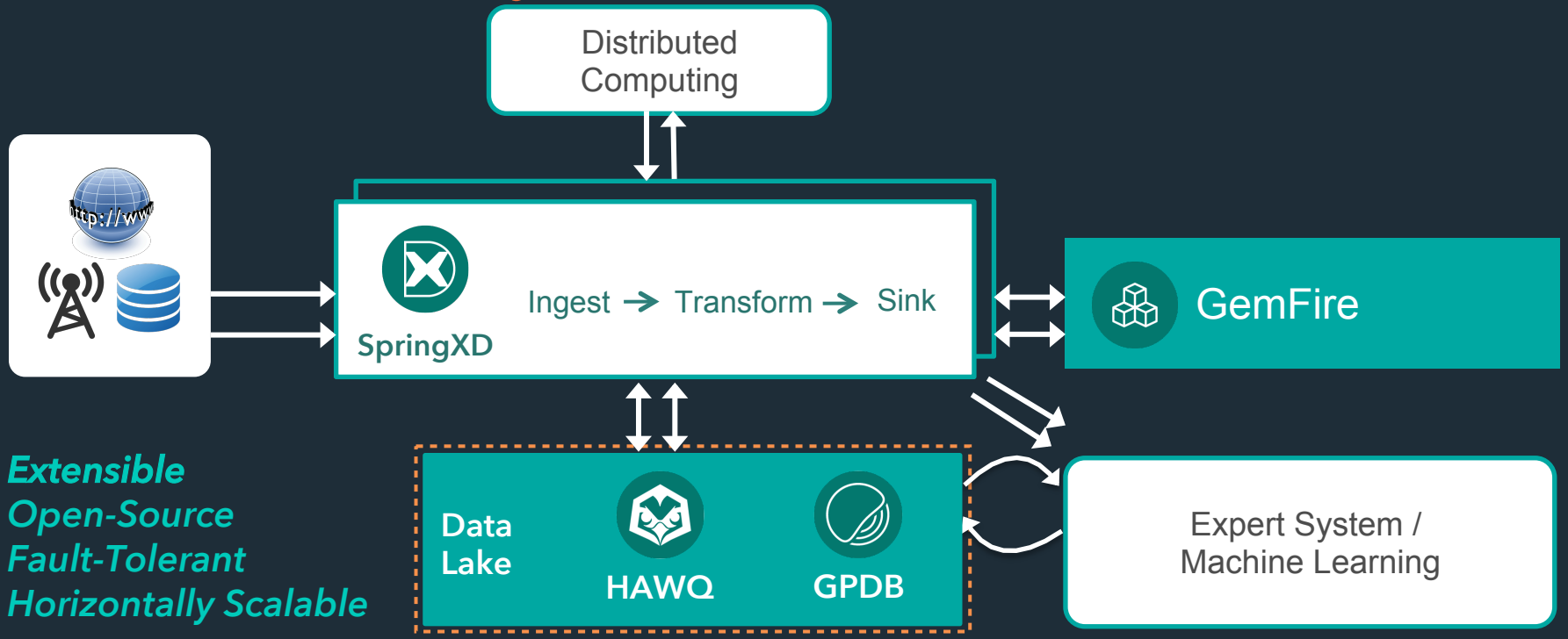
In-memory Real Time Data

- In-Memory Enterprise Data Grid
- Horizontally Scalable, Consistent, Highly Available
- Event handling
- Continuous Queries
- Enterprise Data Geo Distribution



Pivotal

Pivotal Provides SQL Based Advanced Analytics



Extensible
Open-Source
Fault-Tolerant
Horizontally Scalable

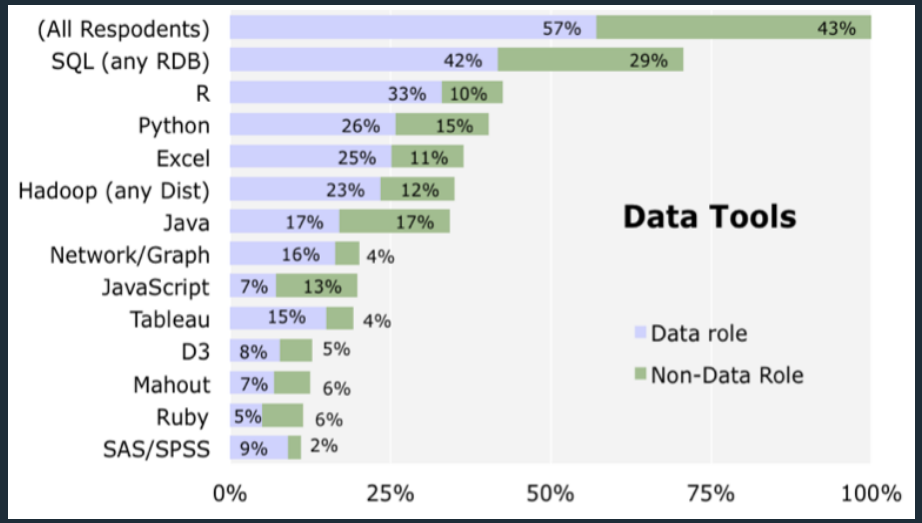
Pivotal

HAWQ

Advanced SQL analytics in Hadoop

- Massively Parallel Processing RDBMS on HADOOP
- ANSI SQL on Hadoop
- Extremely high performance for analytics (not like Hive)
- Stores all data directly on HDFS
- Open-Source

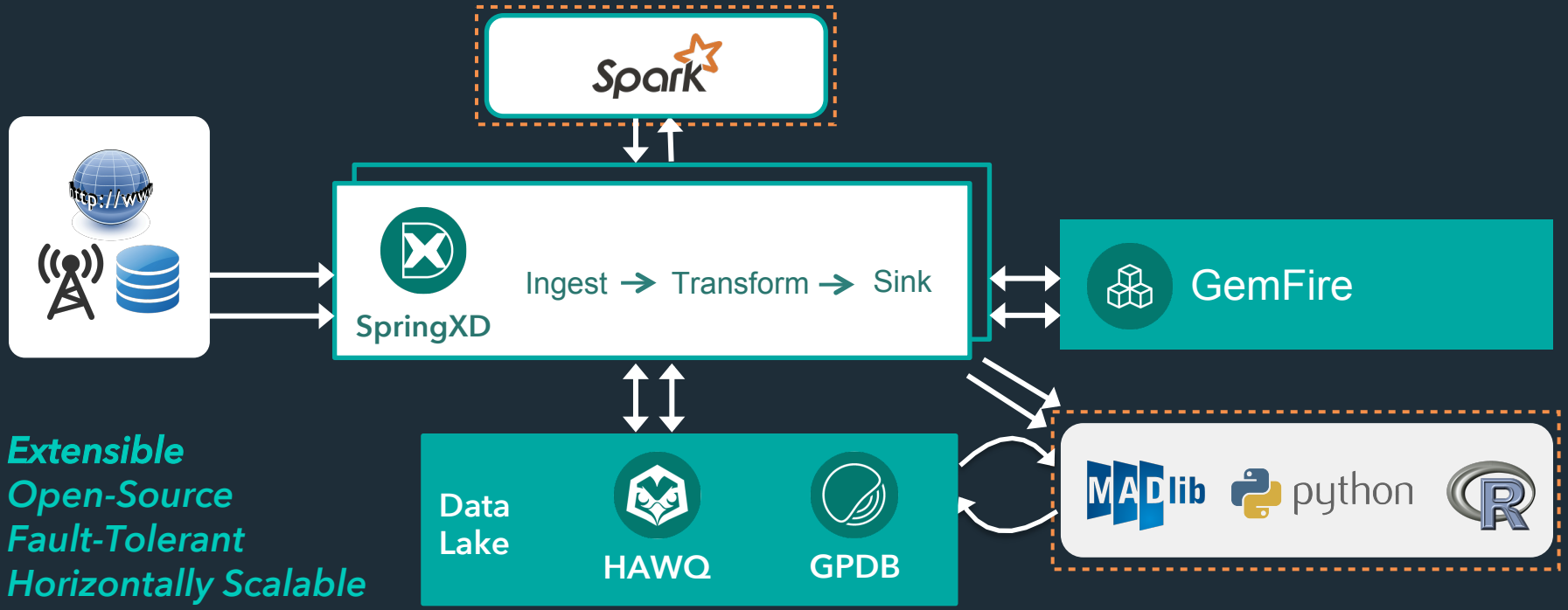
SQL remains #1 choice for Data Science



Combining SQL with Hadoop is key for analytics

Pivotal

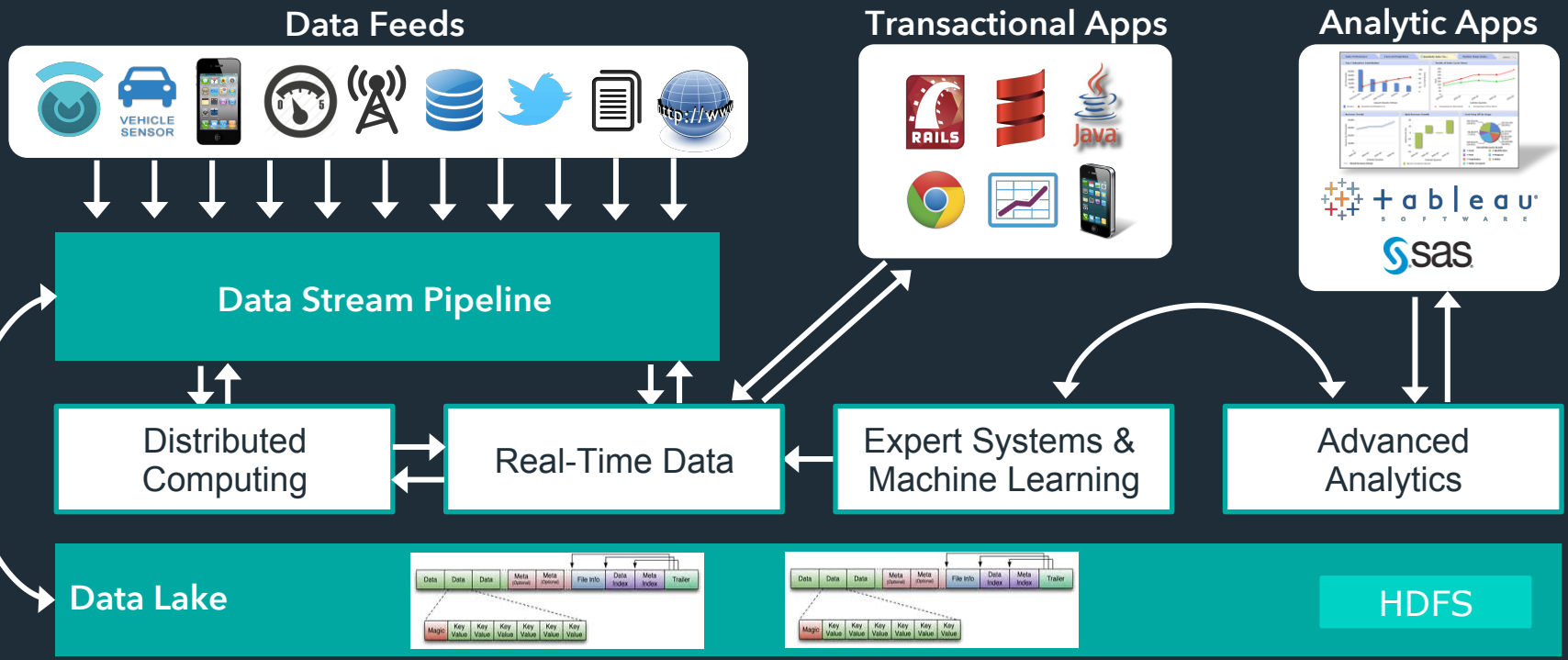
Developers and Data Scientists Can Focus on the Business Value of Data



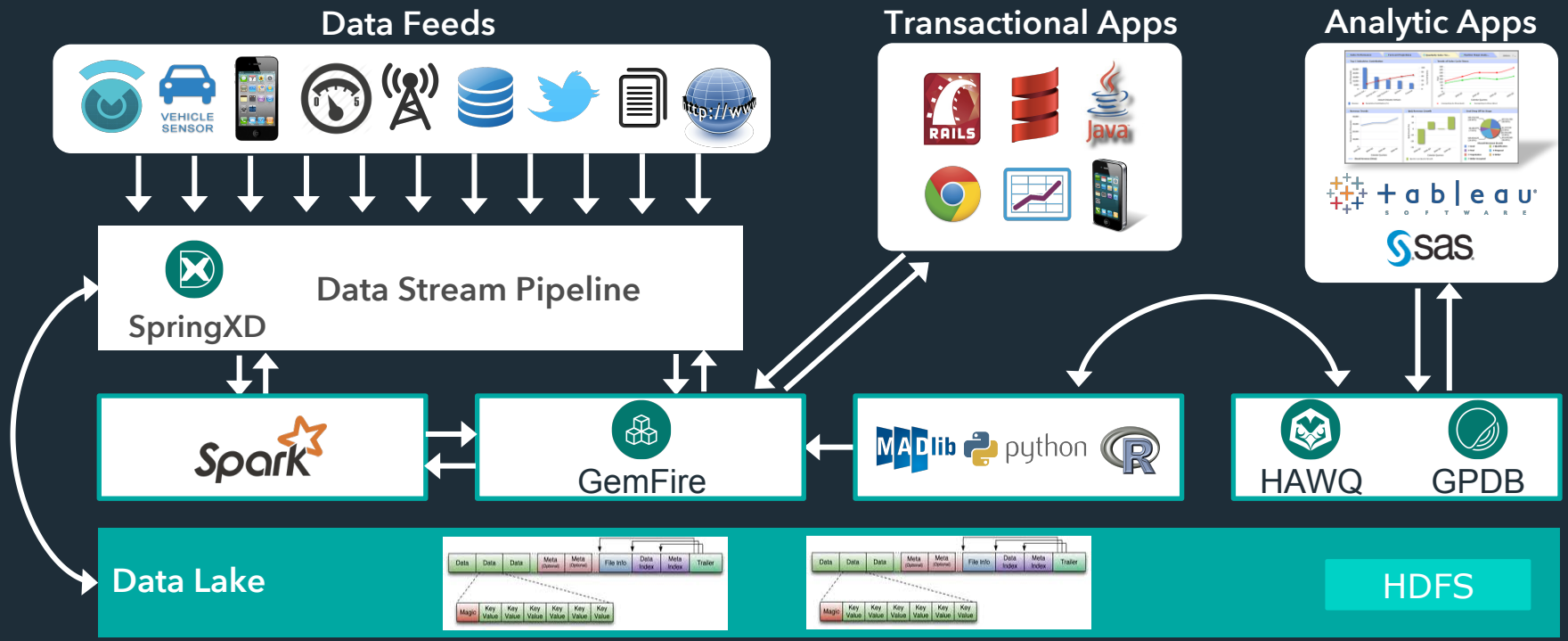
Extensible
Open-Source
Fault-Tolerant
Horizontally Scalable

Pivotal

Data Streaming Reference Architecture



Data Streaming Reference Architecture



Pivotal

“SO WE ARE MOVING TO A WORLD WHERE THE MACHINES WE WORK WITH ARE NOT JUST INTELLIGENT; THEY ARE BRILLIANT.THEY ARE SELF-AWARE, THEY ARE PREDICTIVE, REACTIVE AND SOCIAL. IT'S A WORLD WHERE INFORMATION ITSELF BECOMES INTELLIGENT AND COMES TO US AUTOMATICALLY WHEN WE NEED IT WITHOUT HAVING TO LOOK FOR IT.”

MARCO ANNUNZIATA, GE

The IoT Market Verticals

Diversified Industrial Manufacturers

Agriculture, Security, Retail

Auto Manufacturers

Media (via Mobile devices)

Urban Infrastructure, Cities

Consumer, Connected Home etc.

Healthcare, Life Sciences

Pivotal


“
THE MAGIC HAPPENS WHEN YOU MARRY THE TRADITIONAL ENGINEERING APPROACH WITH THE DATA SCIENCE ENABLED BY THE DATA LAKE. IT OPENS UP A WHOLE NEW WORLD OF POSSIBLE ‘WHAT IF’ QUESTIONS.
”
DAVE BARTLETT, GE AVIATION

GE Aviation – Big Data & IoT

- Goal
  - Improve jet engine efficiency and increase service profitability
  - Unable to store & analyze massive amounts of data for analytics
- Solution
  - **LARGE DATA SETS** ingested via batch
  - Store 100s TB of engine data in Hadoop (PHD)
  - Open doors for industrial engineers to poke at data (HAWQ)
  - **FAST MACHINE LEARNING** based algorithms
    - 2000x faster, 10x cheaper
  - Customer portals for visibility

Pivotal


“
THE REAL OPPORTUNITY FOR CHANGE...SURPASSING THE MAGNITUDE OF THE CONSUMER INTERNET...IS THE INDUSTRIAL INTERNET, AN OPEN, GLOBAL NETWORK THAT CONNECTS PEOPLE, DATA AND MACHINES.
”
JEFF IMMELT, CEO, GE

GE Energy – Fast Data & IoT

- Goal
  - Failing gas turbines causing issues with power generation
  - Unable to store & process fire-hose of data
- Solution
  - **HIGH VELOCITY** data ingestion from Gas Turbines
  - Store 10 TB of turbine data in memory (GemFire)
  - **VERY LOW LATENCY** and **HIGH SPEED** data access
  - “Predictive” Maintenance

Pivotal


“
… YOU USE THOSE DEVICES TO INSPECT CARS AND KEEP TRACK OF INSPECTION RECORDS. THAT CAN THEN FLOW INTO ASSET MANAGEMENT SOFTWARE TO PROVIDE PREDICTIVE ANALYSIS OF WHEN THE ASSET NEEDS TO BE MAINTAINED ... WHEN YOU BOIL THAT DOWN TO BUSINESS, IT’S ABOUT COMPETITIVE ADVANTAGE.
”
BRAD HOWELL, LODESTAR LOGISTICS

GE Transportation – Big & Fast Data

- Goal
  - Help rail companies manage locomotives better
  - Fast data from tracks & Big data from sensors in locomotives
- Solutions
  - **MACHINE LEARNING MODEL** built from combined data set (MADLib)
  - **REAL TIME SCORING** of rail sensor data (Spring XD)
  - **REAL-TIME ALERTING** of critical events via email & a **REAL TIME DASHBOARD** (Spring)

Pivotal

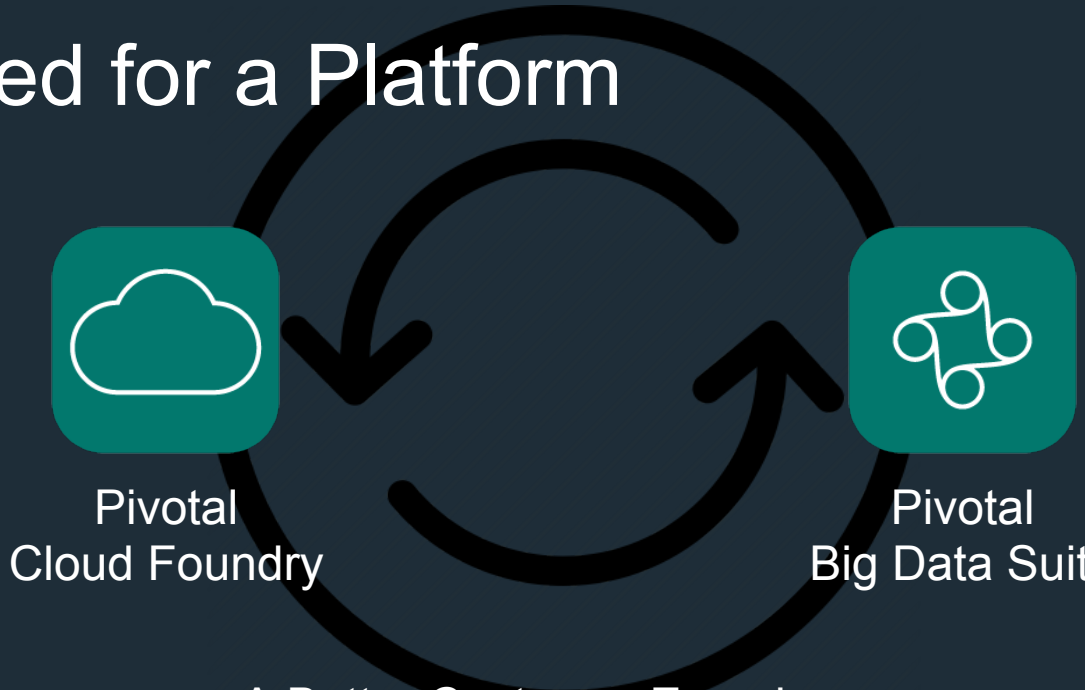

“
WE EXPECT THE PRECISION AGRICULTURE SPACE TO CONTINUE TO GROW QUICKLY AS DATA BECOMES CHEAPER TO STORE AND EASIER TO MOVE FROM PLATFORM TO PLATFORM. WE ARE JUST BEGINNING TO EXPLORE ALL THE VALUE WE CAN CREATE FOR FARMERS WITH THESE TOOLS.
”
BRETT BEGEMANN, MONSANTO

Monsanto – Agriculture & IoT

- Goal
  - Help farmers maximize crop yields
- Solution
  - Use of Big Data to collect & store data from farm equipment
  - Combine with climate and other information
  - Build custom apps using an agile app dev platform (PCF)

Pivotal

IoT - Need for a Platform



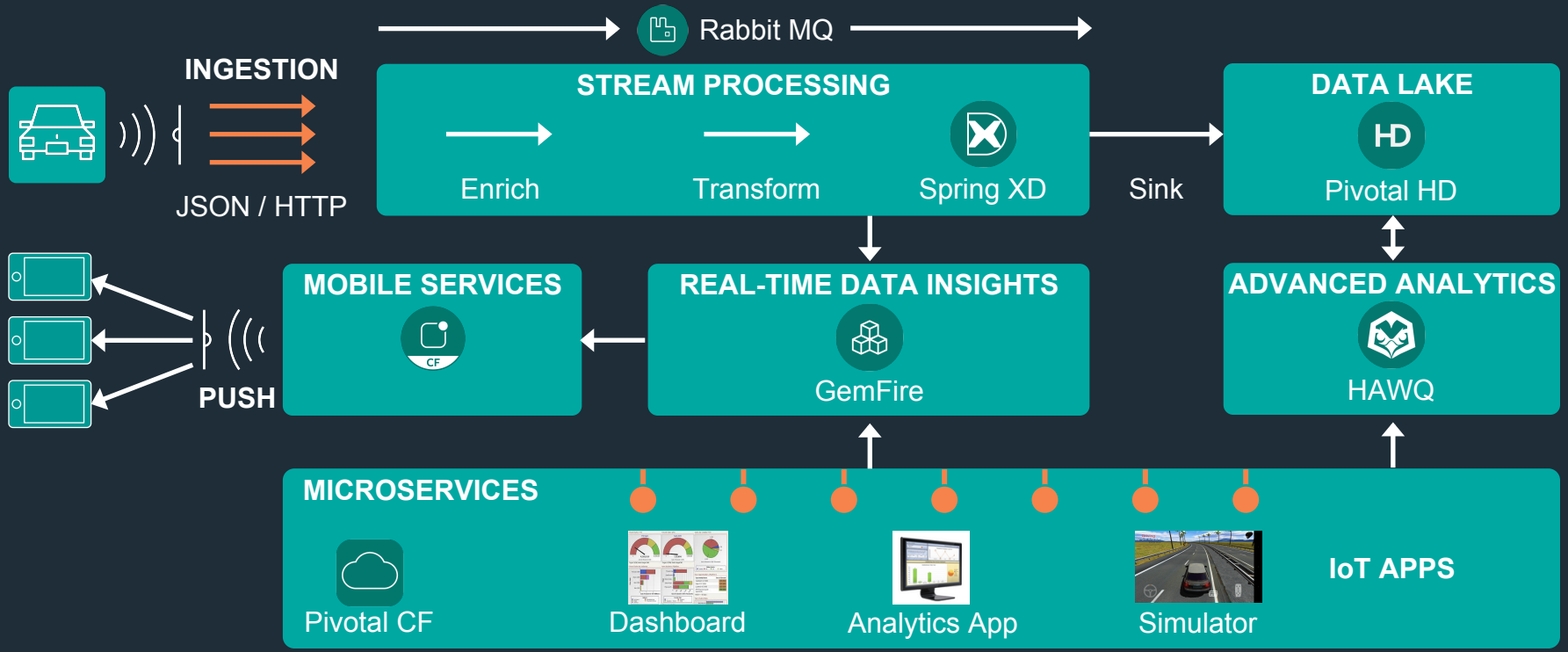
Pivotal
Cloud Foundry

Pivotal
Big Data Suite

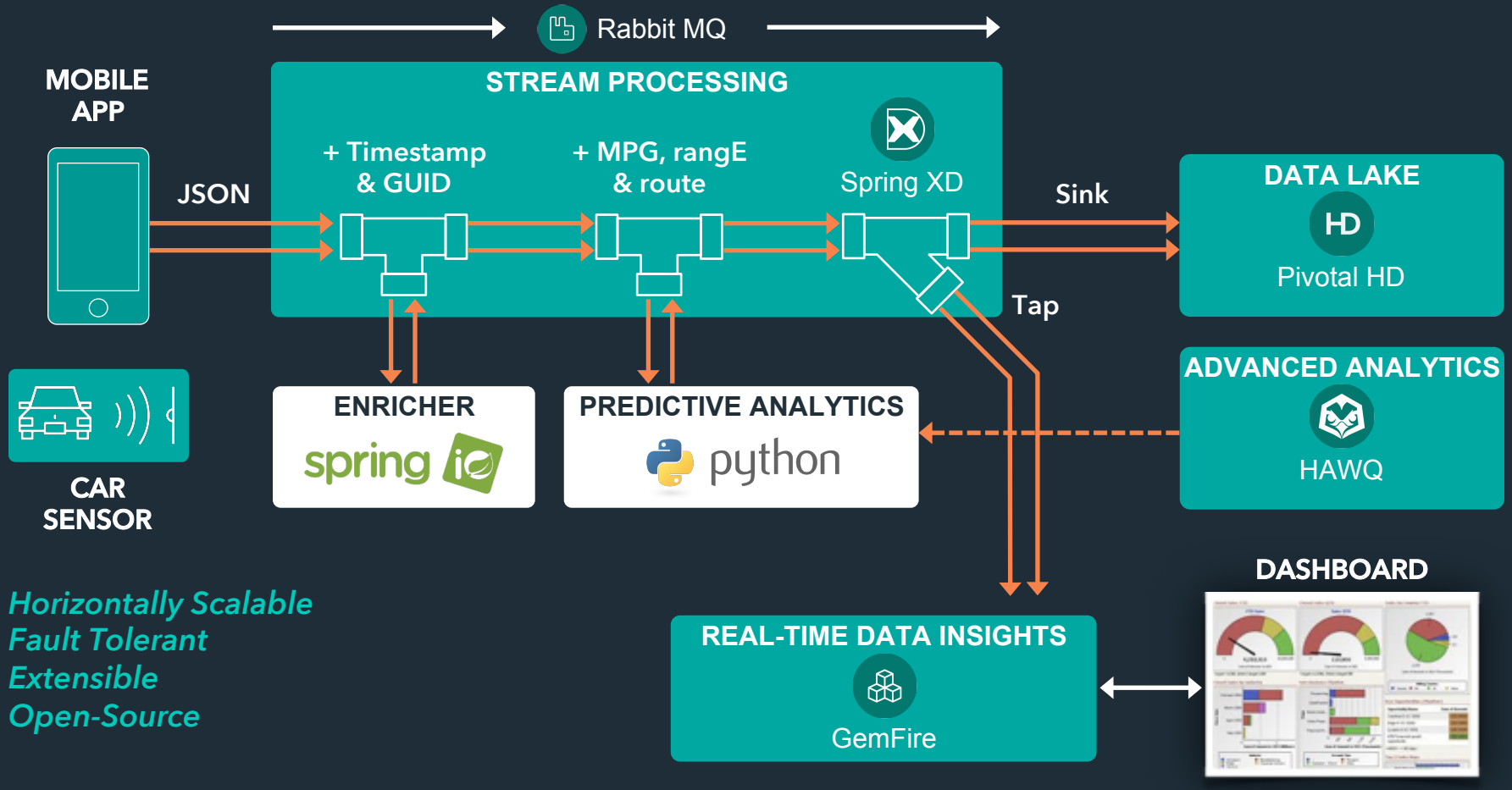
A Better Customer Experience
Using Innovative **Data-Driven Apps**
On a **Integrated Platform**

Pivotal

The Connected Car Architecture



Pivotal



Pivotal

BUILT FOR THE SPEED OF BUSINESS