EXECUTIVE BRIEFING:
The Hidden Data Scientists Lurking in your Company
Putting Things into Perspective

INVESTMENT IN NEXT-GEN VS. LEGACY TECHNOLOGIES FOR DATA

Source: IDC, Gartner; Analysis & Estimates: MapR
Next-gen consists of cloud, big data, software and hardware related expenses
The Impact of AI

McKinsey estimates AI techniques have the potential to create between $3.5T and $5.8T in value annually across nine business functions in 19 industries.

Forrester Research predicts that by 2020, businesses adopting Machine Learning, AI, and Deep Learning, the Internet of Things (IoT), and Big Data will take away more than $1.2 trillion from their less-informed peers.
Demand for Data Scientists

29% annual Increase in Demand
344% since 2013
- Indeed Job Stats

In the UK 80% of companies are planning to hire data scientists
- MHR Analytics
Deep Learning Algorithms

- **Deep Neural Networks**
  - Providing lift for classification and forecasting models

- **Convolutional Neural Networks**
  - Feature extraction and classification of images

- **Recurrent Neural Networks**
  - For sequence of events, language models, time series, etc.
ML/AI ALGORITHMS + DATA + DOMAIN EXPERIENCE
Focus On Machine Learning Tools
Curiosity
The Importance of Data
“90+% of Machine Learning Success is Data Logistics”
More DATA beats complex algorithms

The Unreasonable Effectiveness of Data, published by Google
More Data Allows You to Spot Infrequent Behaviors

With recent data you have limited historical data accumulated
More Data Allows You to Spot Infrequent Behaviors

With big data, you can trace infrequent patterns through time that call out anomalies.
How to see beyond the obvious
How to see beyond the obvious and react quickly
The Importance of Data Logistics

“Machine learning offers a fantastically powerful toolkit for building complex systems quickly. It is remarkably easy to incur massive ongoing maintenance costs at the system level when applying machine learning.”
The Impact of Analytics
Evolving Analytics at Scale

- Descriptive – What happened?
- Predictive – What will happen?
- Prescriptive – The best response to what is happening


Injecting analytics into operations
Digital Growth Requires Complex Requirements

CONTEXT

SPEED

ACTION

NoSQL  Document Database  Analytics  Storage  Processing Engines  Streaming  Messaging
AMERICAN EXPRESS

Implemented big data machine learning use cases: fraud detection & prevention; new customer acquisition & recommendations for better customer experience.

Chao Yuan, SVP & Head of Decision Science

$1Trillion < 2ms

Protected Annually from Fraud

It Takes To Make A Decision
Machine Learning based predictive maintenance & data intensive applications to reduce operational expenses and increase uptime

200K+ Sensors

300 Billion+ Data Points Evaluated Daily
Automation to Aid not Replace Data Scientist
Analytics at Scale: Impact on Business Models

- Moving from upfront payments to usage models
- Paying for the actual value created creates new sources for competitive advantages

Incitec Pivot Ltd
From Manufacturing Optimization to “Explosions as a Service”

Objectives:
• Improve explosive manufacturing quality and yield
• Offer new service to consumers to ensure successful blasts at mine sites

• Collecting data from PLCs and sensors to optimize operations and perform predictive maintenance
• Edge deployments at mine sites for analysis and control of explosions
Training Resources

**Core Platform Services**
- IT/Infrastructure
  - Installation
  - Migrations
  - SLA Plans
  - Best Practices
  - Performance
  - Tuning

**Converged Platform**
- Linux
  - Networking
- Data Center
- Storage
- Operations

**Big Data Workflows**
- BI / DBA
  - Hive/Pig/Spark
  - Oozie/Sqoop
  - Flume
  - MapR-DB/HBase
  - Data Pipeline
  - MapR Streams

**Solution Design**
- Development
  - HBase/MapR-DB
  - Map/Reduce
  - Application Development
  - Integration

**Advanced Analytics**
- Modeler / Analyst
  - Use case Discovery
  - Use case Modeling
  - POC
  - Workshops

**Data Engineering**
- BI / ETL / Reporting
  - Scripting / Java
  - Hadoop MR
  - Eco Projects
  (HBase, Hive, ...)

**Data Science**
- Java
  - Hadoop Developer
  - Architectural Design

**Skills**

© 2019 MapR Technologies, Inc. // MapR Confidential
Quick Start Solutions: Speeding Time-to-Value

- Self Service Data Exploration and BI Analytics on Hadoop
- Data Warehouse Offload, Optimization and Analytics
- Financial Services – Fraud Detection, Anti-Money Laundering
- Retail – Customer 360, Social Media Analysis, Recommendation Engine
- Oil and Gas – Tier 2 Historian
- Complex Event Processing with Drools / Stream Processing
- Time Series Analytics, NoSQL Webstore Applications
- Deep Learning on GPUs for Image Analytics
Data Scientist Access To Fast Moving Innovation & Agility

Required Data

- HDFS API
- NFS File Access
- CSI Container Access
The Challenge for IT Organizations

Innovation
Flexibility
Democratization of Data

Control
Security
Protection
Availability

AI, ANALYTICS

Customer Engagement
Fraud detection
Predictive Maintenance
Anomaly Detection
ML/AI data logistics
BI analytics
Data science platform
Recommendation engine

DATA LAYER IS THE LEVERAGE POINT

IoT, EDGE, CLOUD & CONTAINERIZED

Inter-cloud data/app portability
Cloud bursting
Persistent app in containers
Secure data provisioning
Data modernization
Scale out data lakes
Multi-temperature data
Data catalogs
Analytics at Scale: Importance of Location

- Edge – Single Location Processing
- On Premise – Centralized Processing
- Multi-Cloud – Fully Distributed

➢ Learning Globally, Acting Locally


© 2019 MapR Technologies, Inc. // MapR Confidential
Customers Tell Us They Need A Global Data Fabric

Many Small Sites
A Few Big Sites
Leverages Cloud and On-Premise

Globally Protected
Globally Accessible
Globally Managed

NFS  POSIX  REST  HDFS  Geo-Replication  Data Protection

Self Healing
Assume Failures are Common
Top Issues for Cloud Today

What are the biggest challenges for organizations engaged with public cloud today?

- Security: 66%
- Governance & compliance: 60%
- Staff Lacks Cloud Experience: 58%
- Privacy: 57%
- Vendor lock-in: 47%
- Cost: 40%
- Lack of visibility: 37%
- Unplanned outages: 31%
Cross Cloud: Difficult to Establish Unified & Secure Data Access
Cloud With A Data Platform: Portable, United Access with Security

Application → API Connector

Silo Problem solved!

- Unified Security Model
- Data access decoupled from physical storage location. Globally.
- Data made portable
- No lock-in to proprietary APIs
- Full openness

GLOBAL DATA MANAGEMENT

- Edge
- Private Cloud On Premise
- Public Cloud
- Public Cloud
- Public Cloud

Open APIs
Performs real-time analysis to optimize Oil and Gas drilling and production
Connected Car: Driving the Edge Use Cases

By 2020, more than 250 million vehicles will be connected globally

- Gartner

Data Driver for:

Vehicle Efficiency and Performance
Personalized Experience
Manufacturing Optimization
Driver Assisted and Driverless Vehicles
Requires Ability to Learn Globally and Act Locally

REQUIREMENTS

• Collect data in full fidelity
• Apply the latest, most accurate models – even in areas of low bandwidth, high latency, or space constraints
• Avoid the patchwork of silos that comes with historian systems

COORDINATED DATA FLOWS

• Pub-sub data streaming:
  • Data ingestion from IoT devices and
  • Edge to cloud data communication where streams are persisted for centralized analysis
• Extends the platform to the edge, allowing for applications and models to run at the edge (required for time-sensitive applications)
How to Unleash your Lurking Data Scientists
Machine Learning Logistics

Discrete Response System

Move to Support Multiple Models for accuracy, separation of concerns

Load Balancer Approach

With a load balancer, you can start and stop new models pretty easily, you lack:
• Latency guarantees,
• Ability to compare models on identical inputs,
• Records of all of the requests with responses from all live models.
Stream-based Logistics
Rendezvous Architecture
Abstraction: a Short History of IT

1970

MAINFRAMES
Blackbox

LOCK-IN SPECIALIZATION

CLIENT/SERVER
Specialized HW with open industry software standards (TCP/IP, X86, NFS)

VIRTUAL MACHINES
Software used to abstract Hardware from OS
Freedom to run multiple OS on the same HW

FUNCTION VIRTUALIZATION
Software replaces specialized HW

CONTAINERS
Resources entirely managed in Software

FLEXIBILITY FREEDOM

2019+
Data is a resource as well
Turning Data into a Manageable Resource
All in Software

- Data Containerization
- Global Multi-Tenancy
- Data Portability
- Resource Isolation
- Workload independence
- Security
- Global Web-Scale Deployments
- Performance
- Universal Access

All Managed by Policies in One Layer
Global Data Fabric Supports Existing and New Applications including AI
The Solution for Organizations to Support Analytics

**AI, ANALYTICS**
- Customer Engagement
- Fraud detection
- Predictive Maintenance
- Anomaly Detection
- ML/AI data logistics
- BI analytics
- Data science platform
- Recommendation engine

**DATAWARE FOR DATA-DRIVEN TRANSFORMATION**
- Inter-cloud data/app portability
- Cloud bursting
- Persistent app in containers
- Secure data provisioning
- Data modernization
- Scale out data lakes
- Multi-temperature data
- Data catalogs

**Innovation**
- Flexibility
- Democratization of Data

**Control**
- Security
- Protection
- Availability
Finding and Empowering Your Data Scientists

ML/AI ALGORITHMS + DATA + DOMAIN EXPERIENCE
Meetup Tonight

Online Evaluation of Machine Learning Models
Ted Dunning PhD, MapR CTO

The Microsoft Reactor London
70 Wilson St
Finsbury
EC2A 2DB
United Kingdom

Thursday, May 2, 2019 from 7:00 PM to 8:30 PM (BST)
Rate today’s session

Executive Briefing: The hidden data scientists lurking in your company

Jack Norris (MapR Technologies)
14:05–14:45 Thursday, 2 May 2019
Executive Briefing and best practices, Strata Business Summit
Location: Capital Suite 13
Secondary topics: AI and machine learning in the enterprise

Session page on conference website
Questions?