The Rembrandt Group
Strategies for BIG DATA
2015-2016
Big Data

Interesting applications are data hungry
Increased number & variety of sources
Realization that delete is ‘not an option’
The data grows over time & is immobile
100 TB @ 1Gbps ~= 10 days
Compute needs to go the data
Dramatic decline of hardware cost
Big Data clusters are the new libraries

Total digital data to be created this year **270,000PB** (J. Campbell, et al., Intel Research Pittsburgh, 2007)

Source: Open Cirrus Project
Big Data

Amount of Stored Data By Sector
(in Petabytes, 2009)

Sources:
"Big Data: The Next Frontier for Innovation, Competition and Productivity."

1 zettabyte?
= 1 million petabytes
= 1 billion terabytes
= 1 trillion gigabytes

The traditional ways of managing data fall apart at this scale!
Big Data Problems

• Government
  • Command and Control
  • **Fraud detection and cyber security.**
  • Compliance and regulatory analysis.
  • Content Management
  • Content Archiving
  • Data Protection
  • Record Management
  • Multi-Source Intelligence Analysis
  • Weather Prediction
  • Energy consumption and carbon footprint management.
  • Infectious Disease Detection
  • Crime Prevention
  • Traffic Management
  • Complex Event Analysis and Decision Support
  • **E-Discovery**

• Health & Life Sciences
  • **Fraud Detection**
  • Campaign and sales program optimization.
  • Brand management.
  • Patient care quality and program analysis.
  • Supply-chain management.
  • Drug discovery and development analysis.

• Telecommunications
  • Network performance and optimization.
  • Revenue assurance and price optimization.
  • Customer churn prevention.
  • Campaign management and customer loyalty.
  • Call Detail Record (CDR) analysis.

The ‘big data’ problems span industries and domains and will only get worse
Dimensions of the Problem

Four Dimensions Drive the Need for Big Data Solutions

- Volume
- Variety
- Velocity
- Variability

Source: Forester Research, Hadoop World 2011 Presentation
Two Universes are the New Reality

Source: DeWitt, Nehme, Microsoft Jim Gray Systems Lab

<table>
<thead>
<tr>
<th>Structured data w/ known schema</th>
<th>(Un)(Semi)structured data w/o schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACID</td>
<td>No ACID</td>
</tr>
<tr>
<td>Transactions</td>
<td>No transactions</td>
</tr>
<tr>
<td>SQL</td>
<td>No SQL</td>
</tr>
<tr>
<td>Rigid Consistency Model</td>
<td>Eventual consistency</td>
</tr>
<tr>
<td>ETL</td>
<td>No ETL</td>
</tr>
<tr>
<td>Longer time to insight</td>
<td>Faster time to insight</td>
</tr>
<tr>
<td>Maturity, stability, efficiency</td>
<td>Flexibility</td>
</tr>
</tbody>
</table>

Each solution will have its place in the enterprise.
Why NoSQL?

**SQL:**

- Data Arrives
- cleansing
- ETL
- load
- Analyze data where it lands!

Sometimes termed “Schema First”

**NoSQL:**

- Data Arrives
- Application Program
- SQL System
- Application Program
- SQL Queries

- No cleansing!
- No ETL!
- No load!
- Analyze data where it lands!

Sometimes termed “Schema Later”

Source: DeWitt, Nehme, Microsoft Jim Gray Systems Lab
It all started with Google…

Requirements:
- Scalable to PBs and 1000s of nodes
- Highly fault tolerant
- Simple to program against

Massive amounts of click stream data that had to be stored and analyzed

Two key components:
- Hadoop = HDFS + MapReduce

Source: DeWitt, Nehme, Microsoft Jim Gray Systems Lab
Hadoop

- A storage & processing framework for Big Data
- Open Source version of Google’s work
  - Provides scalability and a high degree of fault tolerance
  - Ability to quickly analyze massive collections of records without forcing data to first be modeled, cleansed, and loaded
  - Easy to use programming paradigm for writing and executing analysis programs that scale to 1000s of nodes and PBs of data
  - Low, up front software and hardware costs
- Think data warehousing for Big Data
- **It does not address the ‘Data integration problem’**
The Data Integration Conundrum

**Situation we have:**

A data-model provides the interface to the data AND the Semantics of the data

For both structured & unstructured data to be integrated it must have a Unified interface

Unified Interface $\Rightarrow$ *Single data-model + Single perspective*

But there is no single right way to represent all Knowledge $\Rightarrow$

There can be no single data-model for all information

**Situation we want:**

A unified interface to All the data that

Does NOT impose a specific Data-model
A Big Data Architecture

- Domain Based and Pre-Packaged Applications; Can be COTS, GOTS or CUSTOM
- Middleware Layer that provides standard, shared application & data services; Provides ‘Platform as a Service’ for Applications
- Infrastructure Layer that provides basic network services, capacity provisioning and system management services; Provides ‘Infrastructure as a Service’ for Platforms

- RemIT Analytics & Visualization Tools
- Data & Processing Syndicate Analytics
- Open Source Application Frameworks
- Data & Processing Syndicate Framework
- UDS Hadoop Distribution (Accumulo, HDFS)
- Open Stack Infrastructure Cloud
- Open Compute-Based Hardware
Big Data Core Components

**1. IMPORT**
Ingest large amount of Structured and Unstructured Data

**2. EXTRACT**
Transform and apply analysis, enrich, reveal meaningful and predictive patterns and extract insights

**3. EXPLORE**
Visualize the data. Predictive dashboards, reports and insights

**Business Intelligence**
Hadoop, Data Marts, Data warehouse, NoSql

**Visualization**
RemIT Big Data Toolkit

- Included with Data & Processing Syndicate (DPS)
  - Syndicate Search Engine using Solr
  - Data-Space NoteBook
  - Data-Space Evolved Video Application
  - Syndicate C2 Sand Table

- Additional RemIT Data Analytics & Visualization
  - Data Ingest and Normalization
  - Text Exploitation w/ NLP
  - CloudBase Geospatial Analysis
  - Visualization and Mapping Widgets
## Use Case: E-Discovery

<table>
<thead>
<tr>
<th><strong>Problem</strong></th>
<th><strong>Solution</strong></th>
<th><strong>Benefits</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Large volume of unstructured and semi-structured data exists, often in disjoint systems.</td>
<td>Big Data Analytics tools with high precision document indexing, search, and retrieval capabilities.</td>
<td>Cheap Commodity Hardware</td>
</tr>
<tr>
<td>Traditional Text Analysis tools cannot handle petabytes and Exabytes of data.</td>
<td>Geospatial Analytics and Visualizations</td>
<td>Open Source technologies with no vendor lock-in</td>
</tr>
<tr>
<td></td>
<td>Semantic and Predictive Analytics</td>
<td>Ability to scale storage and computation as demand grows</td>
</tr>
</tbody>
</table>

Hadoop will be the nucleus of the next-generation EDW in the cloud. – Forrester
Use Case: E-Discovery

1. **Import**
   - Ingest data from various sources

2. **Extract**
   - Transform and apply analysis, enrich and extract insights
   - Create index of feed information, Entity extraction and location information

3. **Explore**
   - Advanced analytics and visualizations to find trends, anomalies, and new relationships
   - Consume vast amounts Twitter feeds,

![Diagram showing data processing system, Lucene/Solr, search, NLP](image-url)
Why Us?

• RemIT’s solution is an Open Source platform with our domain based tools
  • Consists of the Open source Hadoop distribution, Open source Data & Processing Syndicate (DPS) from the Intel community, and RemIT’s own analytics & visualization capabilities
  • Ingests, processes & stores extremely large sets of structured and unstructured data leveraging the Data & Processing Syndicate (DPS)
  • Provides an ecosystem around the data services that includes an application services, analytics, GUI’s and interfaces.

• The Data & Processing Syndicate is a discriminating approach big data
  • Goes beyond the basic ‘big data’ platform
  • Solves fundamental data integration problems
  • An open source platform, in the public domain, and is free to use
  • Has been designed to be simple, fast and applicable to any domain
  • Designed and Implemented in the Intel community with demonstrable benefit

Low Cost, Flexible Approach to Big Data
Summary

• Big Data is not another fad
• Every domain will need to extract knowledge from the growing mass of unstructured and structured data
• There are new tools that can manage the explosion of data in the enterprise
• RemIT offers a unique Open Source approach
• We enable All Data, Information, and Knowledge regardless of Modality, Structure, or Semantics to be Unified within One distributed Data-Space
• Low Cost, Flexible and Domain Oriented solution
• Many deployment options to get started today
Backup Slides
Rembrandt Group
Big Data Core Components

1. IMPORT
   Ingest large amount of Structured and Unstructured Data

2. EXTRACT
   Transform and apply analysis, enrich, reveal meaningful and predictive patterns and extract insights

3. EXPLORE
   Visualize the data. Predictive dashboards, reports and insights

---

**Visualization**

**Business Intelligence**

**Infrastructure**

Hadoop, Data Marts, Data warehouse, NoSql
Big Data Architecture

1. **Import**
   - Ingest data from various sources

2. **Extract**
   - Transform and apply analysis, enrich and extract insights

3. **Explore**
   - Visualize the data

---

**Tools**
- HDFS
- Map Reduce
- Hadoop

**Analytics Apps**
- IT Analyst
- Data Analyst

**Business Intelligence** (Process Data, Semantic and Predictive Analysis)

**BIG DATA**
- Unstructured Data
- Structured Data

**DATA Marts**
- Traditional Data Warehouse (Eg Teradata, Oracle)

**BIG DATA ANALYTICS**
- Traditional BI tools
- BI Analyst

**Unstructured Data**
- Hadoop
- NoSQL DB

**Structured Data**
- HDFS
- Map Reduce

---

**Business Intelligence (Process Data, Semantic and Predictive Analysis)**

**IT Analyst**

**Data Analyst**

**Traditional BI tools**

---

**Import**
- Ingest data from various sources

**Extract**
- Transform and apply analysis, enrich and extract insights

**Explore**
- Visualize the data

---

**Data Marts**
- Traditional Data Warehouse (Eg Teradata, Oracle)
Enables All Data, Information, and Knowledge regardless of Modality, Structure, or Semantics to be Unified within One distributed Data-Space

Allows the Diversity of Exploration, Integration, Enrichment, and Exploitation Processes to be Developed, Tested, and put into Production Together

Provides a Solution to the Data Integration Problem through consideration of data models at a higher level of abstraction

Distill a minimal set of elements sufficient to capture ANY data model
Data & Processing Syndicate

Upper and COI Ontologies

Model DF
(Layer 3)

Harmonized models

Universal store for data/
knowledge models

Source data-models

Data DF
(Layer 2)

Unified data

Universal store for structured data

Artifact DF
(Layer 1)

Standard Interface to Wild

Universal store for unstructured data

Structured Data Ingestion
(e.g. XML)

Indigenous Source Data
(“The Wild”)

Unstructured Artifact Ingestion

Structured Data Ingestion
g. RDB)

External Data Sources & Systems

Enrichment & Exploitation Processing
RemIT Data Analytics & Visualization

- **RemIT Data Ingest and Normalization**
  - Unstructured data analysis and extraction of metadata
  - Categorization, clustering, and taxonomy generation
  - Entity and attribute extraction and normalization
  - Relationship and event coreference

- **Detect and extract text content from documents**
  - HTML, XML, RTF, EPF
  - MS Office & OpenDoc
  - Zip, gzip, and tar
  - Outlook & email archives
  - JPEG and Flash Video
  - Language detection
RemIT Data Analytics & Visualization

- RemIT Text Exploitation
  - Conceptual categorization of objects
  - Parametric based indexing and retrieval
  - Federated retrieval capability
  - Generation of taxonomy information

- Implements text zoning; entity, event, & relationship identification; and document modeling

- Utilizes UIMA, the industry standard, to construct analytic pipelines, implement design patterns, & data model representations

- Natural Language Processing (NLP) – tasks include tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution
RemIT Data Analytics & Visualization

- RemIT Geospatial Analysis

- Big Data Geospatial Indexing, Filtering, Query, store, index, query, analyze, and deliver geospatial raster data

- Hadoop Spatial Query Engine
  - Spatial indexing system based on space-filling curves (Hilbert)
  - Bulk-loading R*-trees with three-dimensional space-filling curves (AoA k-d-tree)

- AOI queries, spatial joins, and spatial analysis operations implemented as MapReduce

- Feature queries performed via CloudBase query language to ease complexity
RemIT Data Analytics & Visualization

- RemIT Visualization Toolkit
- Integrate and enhance analytics as a set of widgets
- Spatially visualize analytic results using lightweight modular OZONE Widget Framework (OWF) widgets
- Expand discovery and search filtering to include points of interest or an area of responsibility – bounding box approach
- Map and layer controls allow operator to easily navigate and display content
- User interaction to select content and display in other widgets via inter-widget communication
Why the Data & Processing Syndicate?

• Up to the platform level they're all pretty much the same
  • All utilize "NoSQL" technologies that turn commodity servers into a distributed database and parallel computation engine
  • They may employ different flavors of "NoSQL" technologies (HBase, MongoDB, Cassandra, ...)
  • We use Cloudbase (NSA GFE) because it provides MLS data security

• The big difference comes in how data is organized and processed
  • NoSQL datastores do not use schemas like relational databases do
  • They simply store data as key-value pairs
  • It’s up to you to figure out how to represent your domain data-model within that structure

• Other DoD & Intel clouds shape the data to suit their analytics
  • Traditional ETL processes transform the data according to some domain-specific data-model
  • Data from disparate sources, even within the same cloud are neither unified nor integrated except by traditional data-model harmonization
  • Which do not scale and often distort / lose information

• The Data & Processing Syndicate breaks the data barriers
  • Provides a unified interface to all data
  • Allows all data to be discovered, explored, searched, enriched, managed, and exploited
  • Semantics provide meaning, not boundaries

• The Data & Processing Syndicate Makes New Things Possible
  • Any number of semantic perspectives may be supported
  • Data can be continuously enriched and utilized within and across multiple contexts
  • Applications can mash-up any data / information / knowledge without data transformations / data-model harmonization
  • Domain-specific data-models can be harmonized in new ways
Our Products & Services

- Big Data Integration
- Big Data Analytics
- Big Data SaaS
- Big Data Appliance
Products & Services

Big Data Integration

- Integration Service Offering
- 1 Month Base Service
- Install Hadoop
- Install Data and Processing Syndicate
- Load a Customer Data sets
- Demonstrate Ingest, Search & Extract
- Basic 2 day technical training
- Customer provides Infrastructure

- RemIT Analytics & Visualization Tools
- Data & Processing Syndicate Frameworks
- Data & Processing Syndicate Analytics

- DPS Hadoop Distribution (CloudBase, HDFS)

- Public API
- Open Stack Infrastructure Cloud
- Open Compute-Based Hardware
Products & Services

Big Data Analytics

- Software Based Offering
- Entire RemIT Visualization, Interface & Analytic Catalog
- Includes Open Source Application Frameworks
- One price for everything plus a yearly maintenance fee for SW updates & support
- Customer must provide infrastructure and Hadoop based platform
- 2 day training included in price
Products & Services

Big Data Appliance

- Product Based Offering
- Provide Entire RemIT Visualization, Interface & Analytic Catalog
- Hadoop Distribution and IMI’s Data Management Services
- The Open Stack distribution for cluster management
- RemIT pre-integrated infrastructure services
- 2 day training, hardware warranty & software maintenance included in the price

- RemIT Analytics & Visualization Tools
- Data & Processing Syndicate Analytics
- Open Source Application Frameworks
- Data & Processing Syndicate Framework
- DPS Hadoop Distribution (CloudBase, HDFS)
- Open Stack Infrastructure Cloud
- Open Compute-Based Hardware
Products & Services

Big Data SaaS

- Service Based Offering
- Big Data Services executed at RemIT datacenters
- Simple per GB price for all services
- Data loaded by customer into RemIT cloud
- Services would be accessible via RemIT portal
- Customer can leave at anytime
- 2 Day training services