What is Big Data?

- Large volumes of complex and variable data that require advanced techniques and technologies to enable capture, storage, distribution, management, and analysis.
- Rapidly expanding volume of high velocity, complex, and diverse types of data.
### Characteristics of Big Data

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Description</th>
<th>Drivers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Volume</strong></td>
<td>Sheer amount of data generated or data intensity that must be ingested, analyzed, and managed to make decisions</td>
<td>Increase in data sources and higher resolution sensors</td>
</tr>
</tbody>
</table>
| **Velocity**  | How fast data is being produced and changed and the speed with which data must be received, understood, and processed | Increase in data sources  
- Improved thru-put connectivity  
- Enhanced computing power of data generating devices |
| **Variety**   | Rise of information coming from new internal and external sources. Structured, Unstructured, Semi-Structured data. |  
- Mobile  
- Social Media  
- Videos  
- Chat  
- Genomics  
- Sensors |

- Strategy calls for USG to “unlock the power of government data to spur innovation across our nation and improve the quality of services for the American people.”
  

- Big Data: Seizing Opportunities, Preserving Values
  
  - White House review of the impact big data technologies will have on a range of economic, social, and government activities

  https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf
Cloud Computing can provide better performance and scalability for most big data systems because they can provide auto-scaling and auto-provisioning.

<table>
<thead>
<tr>
<th>Cloud Characteristic</th>
<th>Big Data Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapid elasticity and scalability</td>
<td>Allows IT services to scale automatically to meet expanding demand for Big Data analytics services and volumes</td>
</tr>
<tr>
<td>Measured service</td>
<td>Big data cloud resources are monitored and controlled per use</td>
</tr>
<tr>
<td>Broad Network Access</td>
<td>Big data cloud resources can be accessed by diverse client platforms across the network</td>
</tr>
<tr>
<td>Resource Pooling</td>
<td>Aggregated Big Data cloud resources in a location-independent manner, enabling them to be assigned and reassigned on demand</td>
</tr>
</tbody>
</table>
How is Big Data Related to IoT?

- IoT consists of internet-connected sensors attached ‘things’, generating and transferring data over a network without requiring human-to-human or human-to-computer

- Increase in connected devices will lead to an exponential increase in the data that needs to be managed

- Big Data capacity is a prerequisite to tapping into the Internet of Things
Overview of the Big Data Market

Data Warehousing/ETL/Data Integration

BI/Visualization/Analytics

Big Data Analytics
## Big Challenges

Percentage of surveyed IT and business leaders listing each of these factors as one of their organization’s top hurdles or challenges with big data:

- Determining **how to get value from big data** - 65%
- Risk and **governance issues (security, privacy, data quality)** - 32%
- Obtaining **skills and capabilities needed** - 30%
- Integrating **multiple data sources** - 30%
- Integrating **big data technology with existing infrastructure** - 29%
- **Defining our strategy** - 28%
- Funding for **big-data related initiatives** - 25%

*Source: Gartner Inc. survey of 302 IT and business leaders in June 2014*
Roadmap
Government is a Data-Driven:

• Each Ministry regularly captures data in non-standardized formats, stored in disparate systems. Scarce resources are spent to maintain this practice and assess this data in an isolated manner – producing only a partial results.
• Citizens capture important information via social media and mobile apps.

This Issue Is:

There is no practical way to aggregate this data and leverage it to bring practical results benefitting Government and Citizens as a whole.
• Accelerating productivity is key to increasing the standards of living.
• Economies as a whole need to enable organizations to take advantage of big data.
Big Data Concept

Standard Tools for Analysis and Visualization

Common Government Big Data Platform

Citizen → Environment → Health → Geospatial → Census

Data Sources
Federal
Local
Citizens
## What it Is

A framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

## Components

<table>
<thead>
<tr>
<th>Component</th>
<th>Component Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadoop Distributed File System (HDFS)</td>
<td>Storage clusters that hold the actual data.</td>
</tr>
<tr>
<td>MapReduce</td>
<td>Java-based system used to process data. Does not involve queries. MapReduce runs as a series of jobs, with each job essentially a separate Java application that pulls out information as needed.</td>
</tr>
</tbody>
</table>
## Notional Data Flow

<table>
<thead>
<tr>
<th>Source &amp; Data Applications</th>
<th>Data Preparation</th>
<th>Data Transformation Metadata Repository</th>
<th>Business Intelligence Decision Support</th>
<th>Analysts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Streaming Data</td>
<td></td>
<td></td>
<td></td>
<td>Industry Domain Expert</td>
</tr>
<tr>
<td>Text Data</td>
<td></td>
<td></td>
<td>Data Mining &amp; Statistics</td>
<td>Analystics Solution End User</td>
</tr>
<tr>
<td>Multi-Dimensional</td>
<td></td>
<td></td>
<td>Optimization &amp; Simulation</td>
<td>Other Analysts and Users</td>
</tr>
<tr>
<td>Time Series</td>
<td></td>
<td></td>
<td>Semantic Analysis</td>
<td></td>
</tr>
<tr>
<td>Geo Spatial</td>
<td></td>
<td></td>
<td>Fuzzy Matching</td>
<td></td>
</tr>
<tr>
<td>Video &amp; Image</td>
<td></td>
<td></td>
<td>Network Algorithms</td>
<td></td>
</tr>
<tr>
<td>Relational</td>
<td></td>
<td></td>
<td>New Algorithms</td>
<td></td>
</tr>
<tr>
<td>Social Network</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Security, Governance, Privacy, Risk Management**

- Data Acquisition
- Filtering, Cleansing, and Validation
- Storage, Hadoop, & Warehousing
- Core Analytics
- Users
Implementation Roadmap

Define Strategy and Concept
Define Governance Framework
Define Use Case
Identify Big Data Assets
Identify Big Data Sources
Assess Data Quality

Phased Implementation

Consistent Communications, Training, Change Management

IT & Security Requirements
Select Technologies
Deploy
Refine Affected Business Processes
Deploy New Business Processes
Tune Applications

Consistent Communications, Training, Change Management
Lessons Learned

- Big Data implementation is iterative and cyclical, versus revolutionary.
- Do not start with technology focus. Instead, consider business/mission requirements that you are unable to address using traditional approaches.
- Address the initial set of use cases by augmenting current IT investments with an intent to scale them to support future deployments.
- Focus on one Big Data “entry point”—volume, variety, and velocity.
- After initial deployment, expand to adjacent use cases, building out a more robust and unified set of core technical capabilities. Examples include the:
  - Ability to analyze streaming data in real time.
  - Use of Hadoop or Hadoop-like technologies to tap huge, distributed data sources.
  - Adoption of advanced data warehousing and data mining software.
• Establish a partnership between industry, academia, and professional organizations to advance big data analytics and maintain professional competency standards.

• Explore which data assets can be made public to help spur innovation outside of Government.
## Additional Resources

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The Cloud Security Alliance</strong></td>
<td>Helping to identify scalable techniques for data-centric security and privacy challenges in big data</td>
</tr>
<tr>
<td><a href="https://cloudsecurityalliance.org/research/big-data">https://cloudsecurityalliance.org/research/big-data</a></td>
<td></td>
</tr>
<tr>
<td><strong>The Open Data Foundation</strong></td>
<td>Helping to promote adoption of global metadata standards and to develop open source frameworks for statistical data</td>
</tr>
<tr>
<td><a href="www.opendatafoundation.org">www.opendatafoundation.org</a></td>
<td></td>
</tr>
<tr>
<td><strong>Apache</strong></td>
<td>Offers a open source library that is a standards-based framework for processing large data sets across clusters of computers</td>
</tr>
<tr>
<td><a href="http://Hadoop.apache.org">http://Hadoop.apache.org</a></td>
<td></td>
</tr>
<tr>
<td><strong>Organization for the Advancement of Structured Information Standards</strong></td>
<td>Will focus on the creation of big data standards.</td>
</tr>
<tr>
<td><a href="www.oasis-open.org">www.oasis-open.org</a></td>
<td></td>
</tr>
</tbody>
</table>
Case Studies
• **Goal:** To develop a tailored intervention approach for different segments of unemployed workers.

• **Action:** Aggregated and analyzed historical data, such as unemployed worker history, the interventions for each worker, and the outcomes.

• **Results:**
  - Identified programs that are ineffective (for improvement or elimination).
  - Refined ability evaluate the characteristics of unemployed and partially employed workers. Developed a segmented approach to offer more targeted placement and counseling services.
  - Over 3 years, reduced spending by €10 billion ($14.9 billion) annually
  - Decreased the amount of time required for unemployed workers to find employment.
Goal: Increase border security around the Great Lakes region between the USA and Canada by enhancing information sharing among public safety agencies.

Issue: Each agency had its own procedures and technology. Data was fragmented among different software systems, databases, spreadsheets, documents.

Solution: Visual Fusion software from IDV Solutions. Provides a way to connect diverse systems and data in a shared, web-based view.

- Multiple agencies use the solution (called REsILlENT) to share information, including 911 incident reports, real-time webcam feeds, regionally available resources, and vulnerable infrastructure, etc.
- Has a “Digital Whiteboard” which allows emergency planners from multiple agencies to visualize possible scenarios and work together on potential responses.
Chicago’s Smart Data Platform

- An open source predictive analytics platform. Connected to WindyGrid, a hub housing information from every department in real time and gathering about 7 million rows of data per day.

- Scalable design and user-friendly interface

- Predictive power of the tool is its ability to analyze data relationships at a speed and on a scale not previously possible.

- 3rd Party Apps are expanding rapidly
  - **Purple Binder** aggregates social services information so that social workers and healthcare professionals have an up-to-date, single source of data about services available to their clients.

*Free to any city willing to install it.*