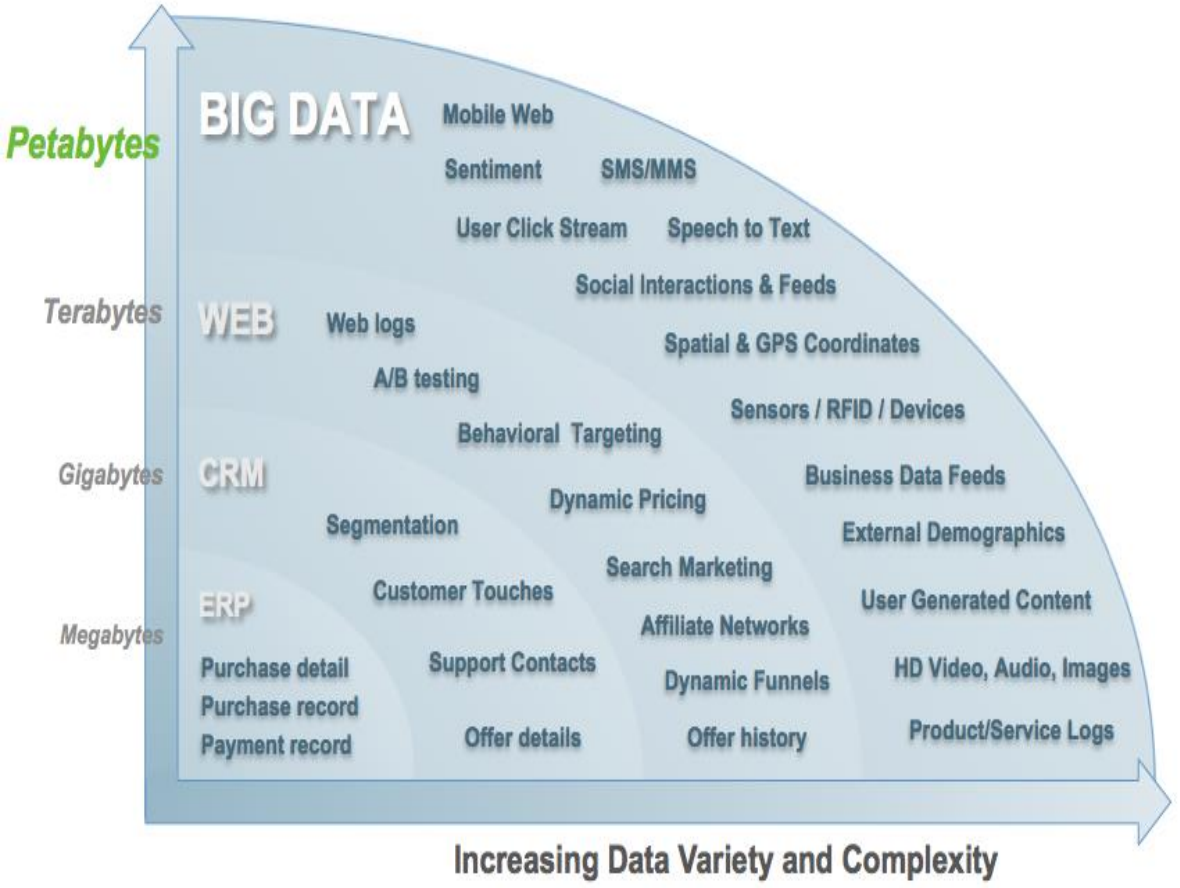
Big Data: Overview and Roadmap

eGlobalTech
Delivering Value, Achieving Results

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

Characteristic	Description	Drivers
Volume	Sheer amount of data generated or data intensity that must be ingested, analyzed, and managed to make decisions	Increase in data sources and higher resolution sensors
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

Big Data in USG

- **"Digital Government: Build a 21st Century Platform to Better Serve The American People" (Digital Government Strategy).**
  - 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."

https://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government.html

- **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

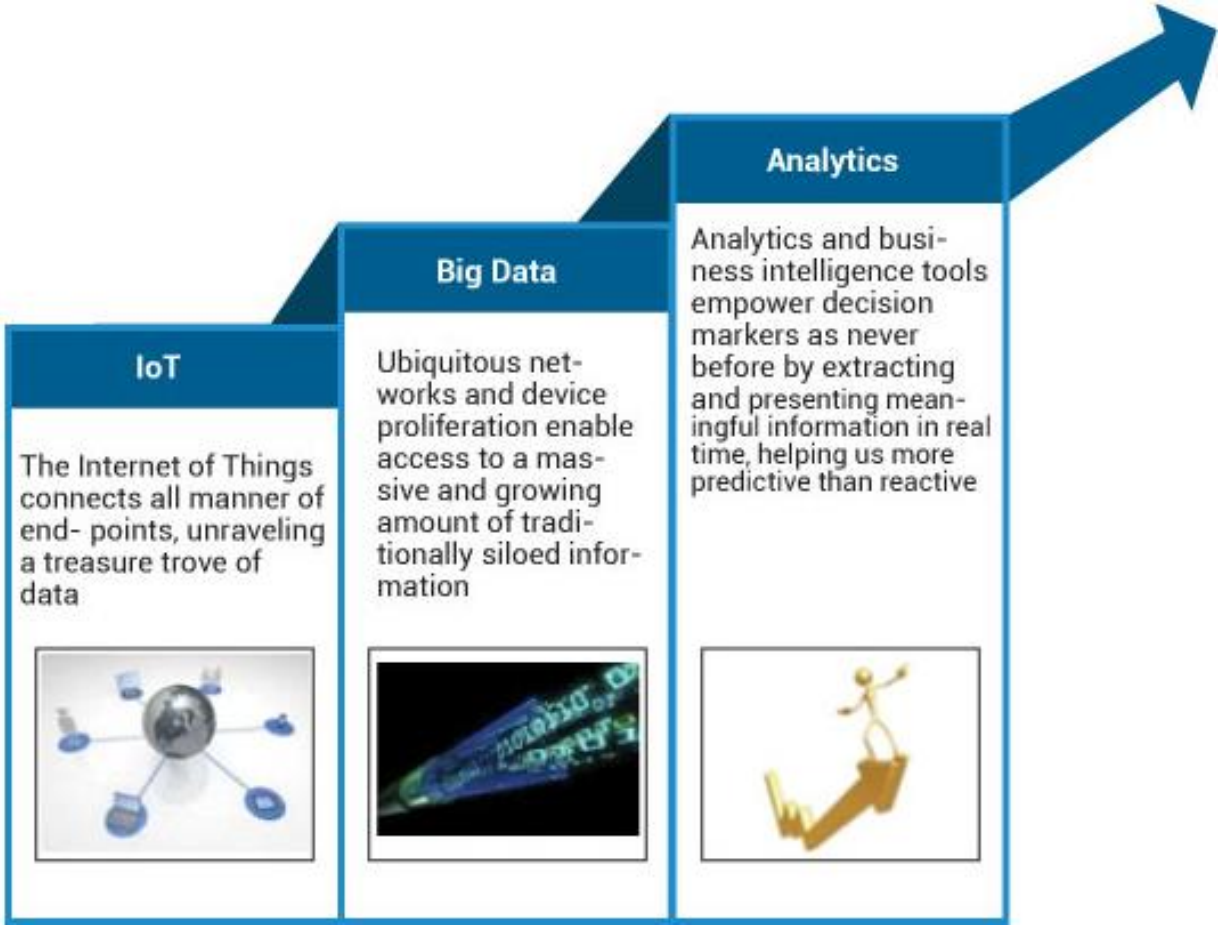
How is Big Data Related to Cloud?

Cloud Computing can provide better performance and scalability for most big data systems because they can provide auto-scaling and auto-provisioning.

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

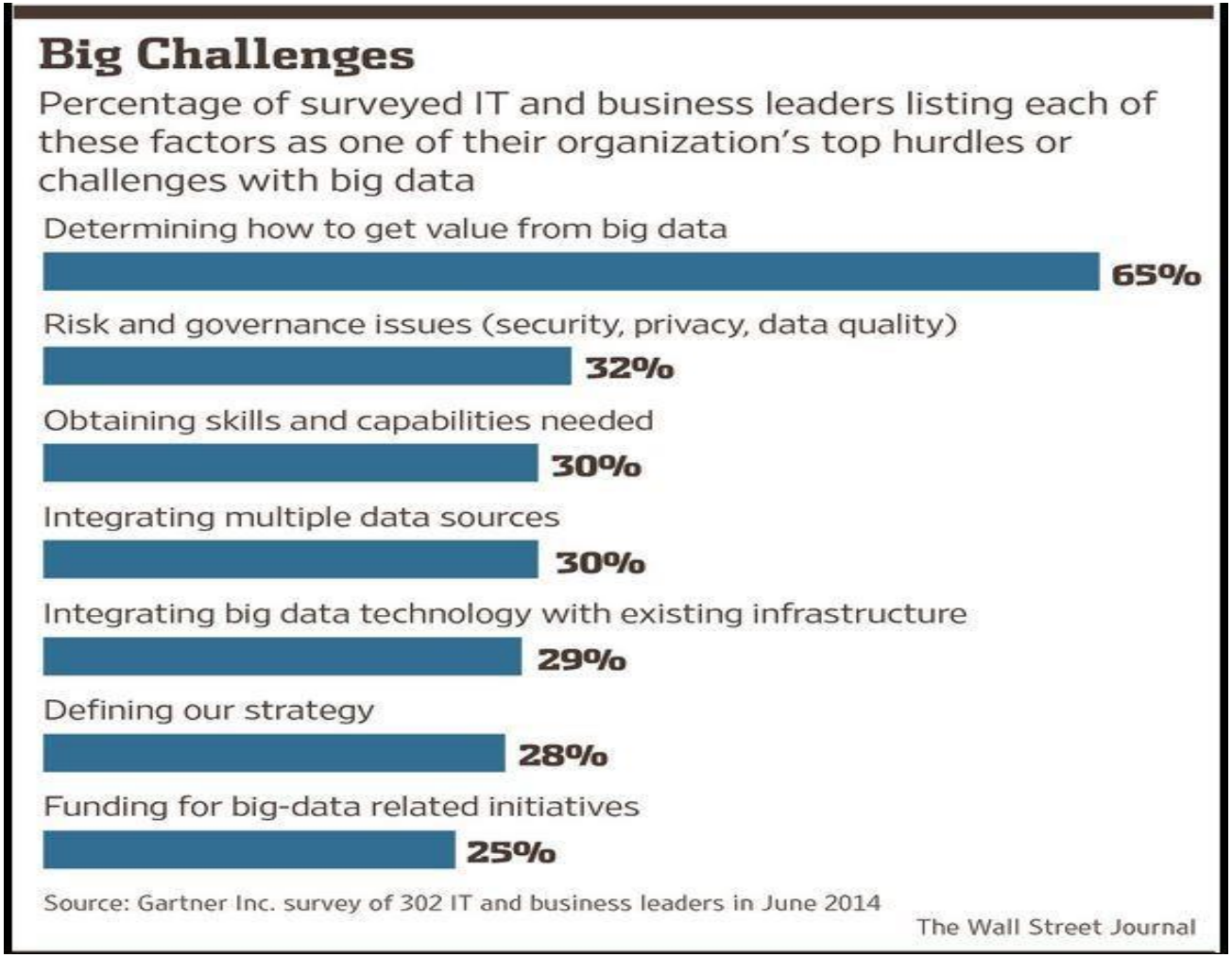
How is Big Data Related to IoT?

- 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

Implementation Challenges



Roadmap

Why Big Data in Government?

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 a 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.

The Business Case

- Accelerating productivity is key to increasing the standards of living
- Economies as a whole need to enable organizations to take advantage of big data.

US health care
- $300 billion value per year
- ~0.7 percent annual productivity growth

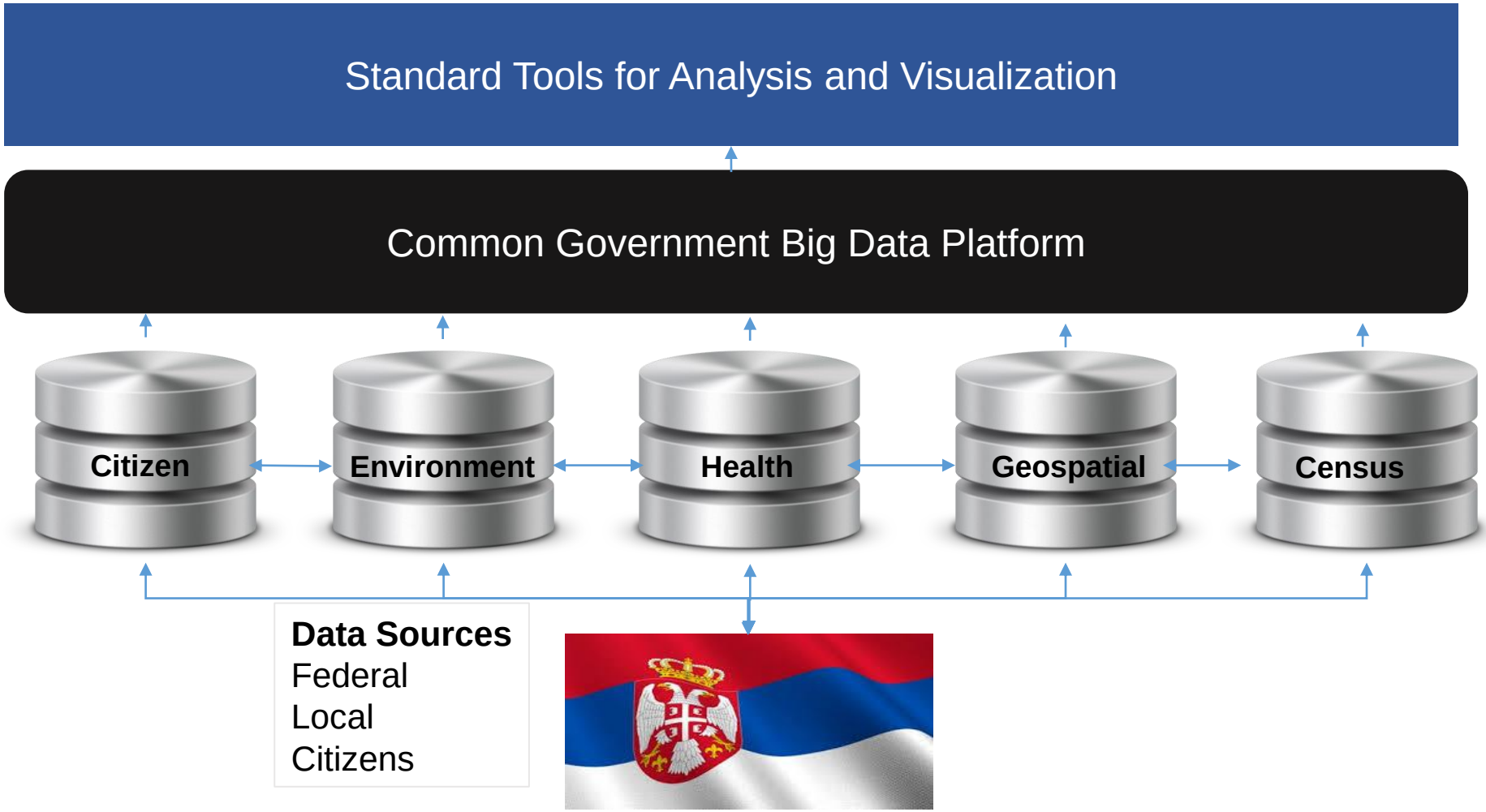
Europe public sector administration
- €250 billion value per year
- ~0.5 percent annual productivity growth

Global personal location data
- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

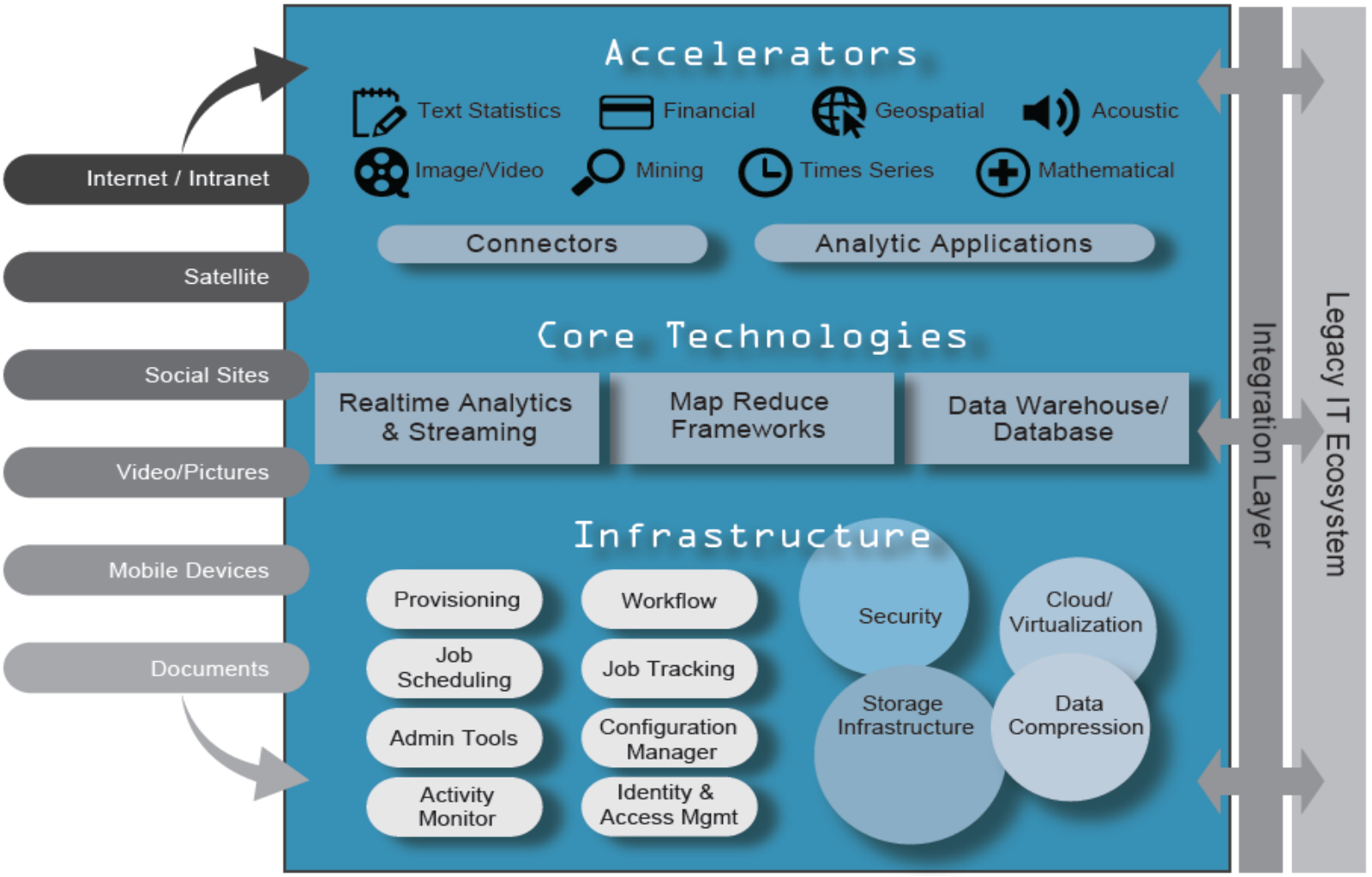
US retail
- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

Manufacturing
- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

Big Data Concept



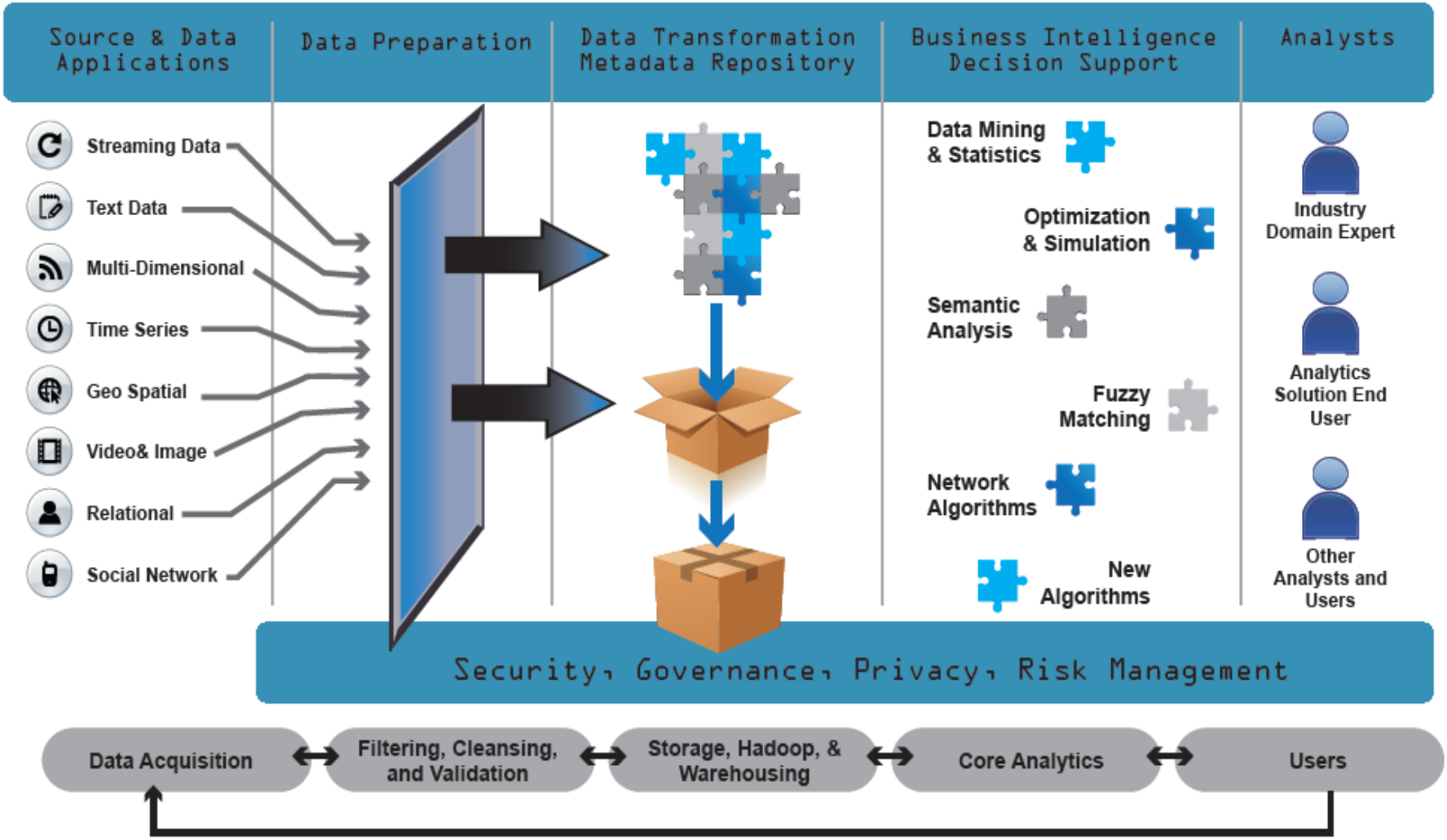
Big Data Enterprise Model



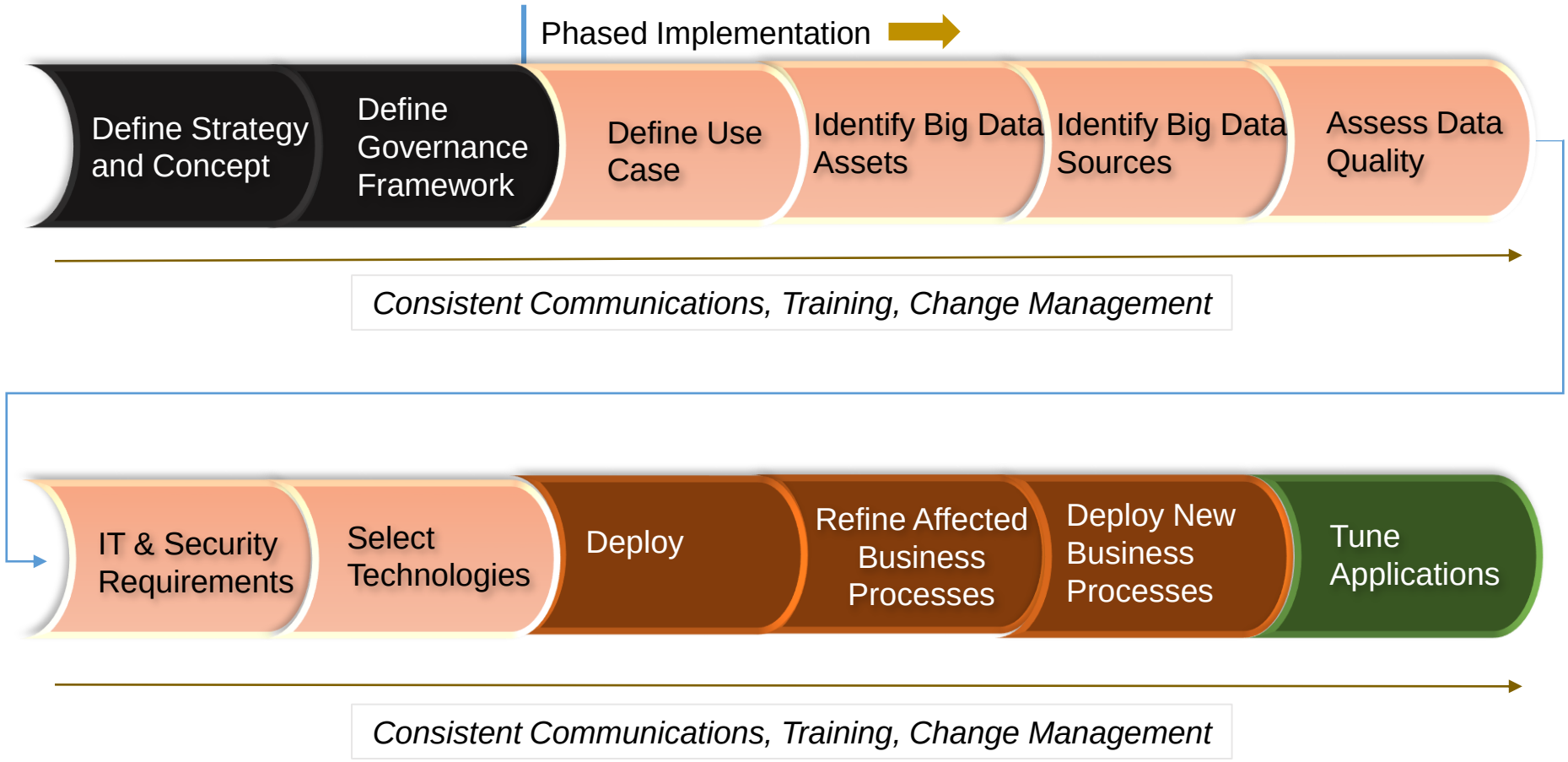
Data Transformation

What it Is	Components	Component Definition
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.	Hadoop Distributed File System (HDFS)	Storage clusters that hold the actual data.
	MapReduce	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.

Notional Data Flow



Implementation Roadmap



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.

Additional Recommendations

- 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

Source	Description
The Cloud Security Alliance https://cloudsecurityalliance.org/research/big-data	Helping to identify scalable techniques for data-centric security and privacy challenges in big data
The Open Data Foundation www.opendatafoundation.org	Helping to promote adoption of global metadata standards and to develop open source frameworks for statistical data
Apache http://Hadoop.apache.org	Offers a open source library that is a standards-based framework for processing large data sets across clusters of computers
Organization for the Advancement of Structured Information Standards www.oasis-open.org	Will focus on the creation of big data standards.

Case Studies

German Federal Labor Agency

- **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.

US Department of Homeland Security

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 REsILIeNT) 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.