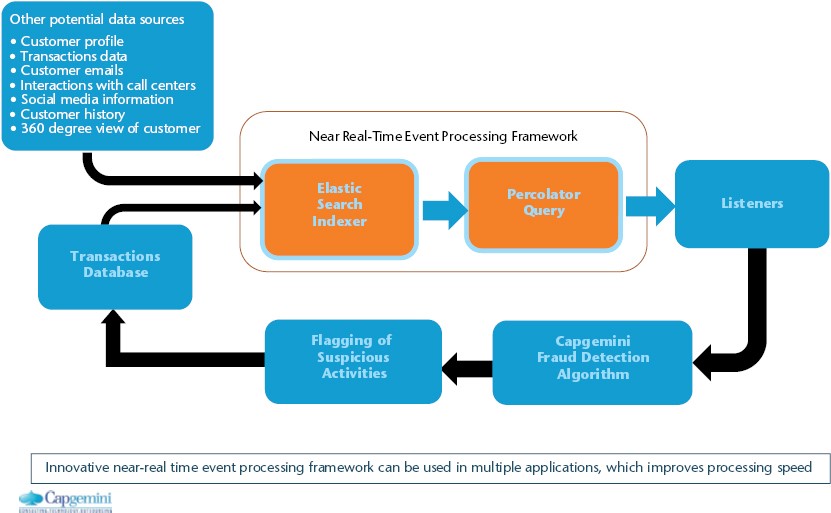
**Fraud and BigData \*\*\***

In order to prevent the fraud, credit card transactions are monitored and checked in near real time. If the checks identify pattern inconsistencies and suspicious activity, the transaction is identified for review and escalation.

The Capgemini Financial Services team believes that due to the nature of data streams and processing required, Big Data technologies provide an optimal technology solution based on the following three Vs:

1. High volume. Years of customer records and transactions (150 billion+ records per year)
2. High velocity. Dynamic transactions and social media information
3. High variety. Social media plus other unstructured data such as customer emails, call center conversations, as well as transactional structured data

Capgemini’s new fraud Big Data initiative focuses on flagging the suspicious credit card transactions to prevent fraud in near real-time via multi-attribute monitoring. Realtime inputs involving transaction data and customers records are monitored via validity checks and detection rules. Pattern recognition is performed against the data to score and weight individual transactions across each of the rules and scoring dimensions. A cumulative score is then calculated for each transaction record and compared against thresholds to decide if the transaction is potentially suspicious or not.



Once the transaction data has been processed, the percolator query then performs the functioning of identifying new transactions that have raised profiles.

**Risk and Big Data \*\***

Many of the world’s top analytics professionals work in risk management. It would be an understatement to say that risk management is data-driven—without advanced data analytics, modern risk management would simply not exist. The two most common types of risk management are credit risk management and market risk management. A third type of risk, operational risk management, isn’t as common as credit and market risk.

The tactics for risk professionals typically include avoiding risk, reducing the negative effect or probability of risk, or accepting some or all of the potential consequences in exchange for a potential upside gain.

*Credit risk analytics* focus on past credit behaviors to predict the likelihood that a borrower will default on any type of debt by failing to make payments which they obligated to do. For example, “Is this person likely to default on their $300,000 mortgage?”

Market risk analytics focus on understanding the likelihood that the value of a portfolio will decrease due to the change in stock prices, interest rates, foreign exchange rates, and commodity prices. For example, “Should we sell this holding if the price drops another 10 percent?”

**Credit Risk Management**

Credit risk management is a critical function that spans a diversity of businesses across a wide range of industries.



Big Data and Algorithmic Trading

**Critical components of Hadoop are:**

* **The Hadoop Distributed File System (HDFS)**. HDFS is the storage system for a Hadoop cluster. When data lands in the cluster, HDFS breaks it into pieces and distributes those pieces among the different servers participating in the cluster. Each server stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server.
* **MapReduce**. Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a comprehensive answer.

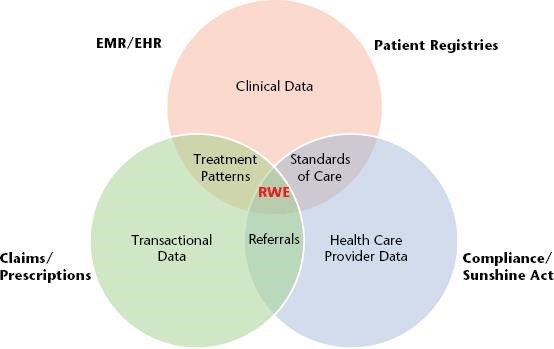
MapReduce is the agent that distributes the work and collects the results.

Both HDFS and MapReduce are designed to continue to work in the face of system failures. HDFS continually monitors the data stored on the cluster. If a server becomes unavailable, a disk drive fails, or data is damaged, whether due to hardware or software problems, HDFS automatically restores the data from one of the known good replicas stored elsewhere on the cluster. Likewise, when an analysis job is running, MapReduce monitors progress of each of the servers participating in the job. If one of them is slow in returning an answer or fails before completing its work, MapReduce automatically starts another instance of that task on another server that has a copy of the data. Because of the way that HDFS and MapReduce work, Hadoop provides scalable, reliable, and fault-tolerant services for data storage and analysis at very low cost.

**Big Data and Advances in Health Care**

Big Data promises an enormous revolution in health care, with important advancements in everything from the management of chronic disease to the delivery of personalized medicine. In addition to saving and improving lives, Big Data has the potential to transform the entire health care system by replacing guesswork and intuition with objective, data-driven science see the following figure

Data in the World of Health Care



The U.S. health care system is increasingly challenged by issues of cost and access to quality care. Payers, producers, and providers are each attempting to realize improved treatment outcomes and effective benefits for patients within a disconnected health care framework. Historically, these health care ecosystem stakeholders tend to work at cross purposes with other members of the health care value chain. High levels of variability and ambiguity across these individual approaches increase costs, reduce overall effectiveness, and impede the performance of the health care system as a whole.

# 3 dimensions / characteristics of Big data

3Vs (volume, variety and velocity) are three defining properties or dimensions of big data. Volume refers to the amount of data, variety refers to the number of types of data and velocity refers to the speed of data processing.

**Volume:**

The size of available data has been growing at an increasing rate.

The volume of data is growing. Experts predict that the volume of data in the world will grow to 25 Zettabytes in 2020. That same phenomenon affects every business – their data is growing at the same exponential rate too.

This applies to companies and to individuals. A text file is a few kilo bytes, a sound file is a few mega bytes while a full length movie is a few giga bytes. More sources of data are added on continuous basis. For companies, in the old days, all data was generated internally by employees.

Peta byte data sets are common these days and Exa byte is not far away.

**Velocity:**

Data is increasingly accelerating the velocity at which it is created and at which it is integrated.

Initially, companies analyzed data using a batch process. One takes a chunk of data, submits a job to the server and waits for delivery of the result. That scheme works when the incoming data rate is slower than the batch-processing rate and when the result is useful despite the delay. With the new sources of data such as social and mobile applications, the batch process breaks down. The data is now streaming into the server in real time, in a continuous fashion and the result is only useful if the delay is very short.

**Variety:**

Variety presents an equally difficult challenge. The growth in data sources has fuelled the growth in data types. In fact, 80% of the world’s data is unstructured. Yet most traditional methods apply analytics only to structured information.

From excel tables and databases, data structure has changed to loose its structure and to add hundreds of formats. Pure text, photo, audio, video, web, GPS data, sensor data, relational data bases, documents, SMS, pdf, flash, etc. One no longer has control over the input data format. Structure can no longer be imposed like in the past in order to keep control over the analysis.

The variety of data sources continues to increase. It includes

■ Internet data (i.e., click stream, social media, social networking links)

■ Primary research (i.e., surveys, experiments, observations)

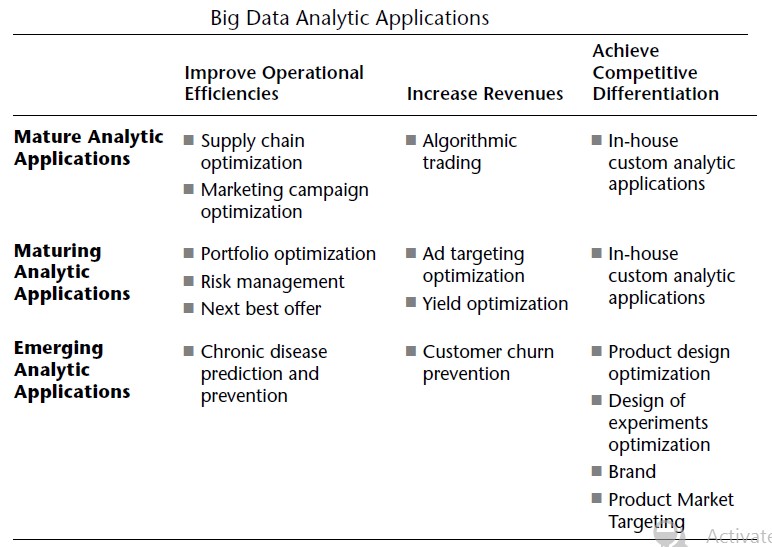
■ Secondary research (i.e., competitive and marketplace data, industry reports, consumer data, business data)

■ Location data (i.e., mobile device data, geospatial data)

■ Image data (i.e., video, satellite image, surveillance)

■ Supply chain data (i.e., EDI, vendor catalogs and pricing, quality information) ■ Device data (i.e., sensors, PLCs, RF devices, LIMs, telemetry)

# Big data analytic spectrum



**Structured vs semistructured vs unstructured**

Structured data generally resides in a relational database, and as a result, it is sometimes called "relational data." This type of data can be easily mapped into pre-designed fields. For example, a database designer may set up fields for phone numbers, zip codes and credit card numbers that accept a certain number of digits. Structured data has been or can be placed in fields like these. By contrast, unstructured data is not relational and doesn't fit into these sorts of pre-defined data models.

In addition to structured and unstructured data, there's also a third category: semistructured data. Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze. Examples of semi-structured data might include XML documents and NoSQL databases.

The term "big data" is closely associated with unstructured data. Big data refers to extremely large datasets that are difficult to analyze with traditional tools. Big data can include both structured and unstructured data, but IDC estimates that 90 percent of big data is unstructured data. Many of the tools designed to analyze big data can handle unstructured data.