Mining Big Data: Its Current Status and Future

Mayushi Chouhan, Rohit Singh Nain
Student, Department of Computer Science, Acropolis Institute of Technology & Research, Indore, Madhya Pradesh, India

DOI: 10.23956/ijarcsse/V7I7/0218

Abstract— Organizations create 2.5 Quintillion bytes of data. So much that 90% of the data in the world today has been set up in the last two years alone. What is Big Data? Big Data is large volumes of structured and unstructured data. This data is what organizations collect on a daily basis. The amount of data is not the important part, but the information gathered from that data is the key. Collecting and analyzing Big Data gives organizations enhanced insight, decision making, and process automation. Approximately each one can agree that big data has taken the business world by storm, but what’s next? Will data continue to grow? What technologies will develop around it? Or will big data become a relic as quickly as the next trend — cognitive technology? Fast data? - appears on the horizon. I believe, am that big data is only going to get bigger and those companies that ignore it will be left further and further behind. This paper studies about what is big data, how does it helps organizations to extract information, its tools and technologies and its future.

Keywords— Big data, decision making, process automation, cognitive technology.

I. INTRODUCTION

What is meant by big data?

Big data is a term for data elements that are so large or intricate that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying and information privacy. Big data is data that's too big for traditional data management to handle. Big, of course, is also subjective.

That's why we'll explain it according to three vectors: volume, velocity, and variety -- the three Vs.

1) Volume

Volume is the V mainly connected with big data because, well, volume can be big. What we're talking about here is quantities of data that arrive at almost incomprehensible proportions.

Facebook, for example, stores photographs. That declaration doesn't begin to boggle the mind until you start to realize that Facebook has more users than China has people. Every user has stored a whole lot of photographs. Facebook is storing roughly 250 billion images.

Can you imagine?

So, in the world of big data, when we start discussion about volume, we're talking about insanely large amounts of data. As we move forward, we're going to have more and more huge collections. Or, consider our new world of connected apps. Everyone is carrying a Smartphone. Let's look at a simple example, a to-do list app. More and more vendors are running app data in the cloud, so users can access their to-do lists across devices. Since many apps use a freemium model, where a free version is used as a loss-leader for a premium version, SaaS-based app vendors tend to have a lot of data to store.

Todolist, for example (the to-do manager I use) has roughly 10 million active installs, according to Android Playstore. That's not counting all the installs on the Web and iOS. Each of those users has lists of items -- and all that data needs to be stored. Todolist is certainly not Facebook scale, but they still store vastly more data than almost any application did even a decade ago.

2) Variety

We have noticed that I've concerning photographs, sensor data, tweets, encrypted packets, and so on. Each of these is very different from each other. This data isn't the old rows and columns and database joins of our forefathers. It's very different from application to application, and much of it is unstructured. That means it doesn't easily fit into fields a database application. Take, for example, email messages. A legal finding process might need sifting through thousands to millions of email messages in a group. Not one of those messages is going to be exactly like another. Each one will consist of a sender's email address, a destination, plus a time stamp. Human-written text and possibly attachments contain in every messages.

Photos and videos and audio recordings and email messages and documents and books and presentations and tweets and ECG strips are all data, but they're typically unstructured, and extremely varied.

Data diversity makes up the variety vector of big data.

Then, of course, there are all the inside enterprise collections of data, ranging from energy industry to healthcare to national security. All of these organizations are generating and capturing vast amounts of data.

That's the volume vector.
3) Velocity
Remember our Facebook example? 250 billion pictures may seem like a lot. But if you want your mind blown, consider this: Facebook users upload more than 900 million photos a day.

Velocity is the parameter of how fast the data is coming in. Face-book has to handle a tsunami of photographs every day. It has to ingest it all, process it, file it, and somehow, later, be able to retrieve it.

Here's another example. Let's say you're running a presidential campaign and you want to know how the folks "out there" are feeling about your candidate right now. How would you do it?

One way would be to license some Twitter data from Gnip (recently acquired by Twitter) to grab a steady stream of tweets, and subject them to sentiment analysis.

**Big data's biggest problem: It's too hard to get the data in**

Even as big data has been turned into more of a marketing term than a technology, it still has huge untapped potential.

That feed of Twitter data is often called "the firehose" because so much data (in the form of tweets) is being formed, it feels like being at the business end of a firehose. Here's another velocity example; packet analysis for cyber security. The Internet sends a huge amount of information across the world every second. For an enterprise IT team, a section of that flood has to travel through firewalls into a corporate network.

Unfortunately, cyber attacks, cybercrime, and cyber espionage rise in day by day scenario, sinister payloads can be able to hidden flow of data passing through the firewall. To prevent compromise, that flow of data has to be investigated and analyzed for anomalies, patterns of behavior that are red flags. This is getting harder as more and more data is protected using encryption. At the very same time, bad guys are hiding their malware payloads inside encrypted packets.

Or take sensor data. The more the Internet of Things takes off, the more connected sensors will be out in the world, transmitting tiny bits of data at a near constant rate. As the number of units increase, so does the flow.

That flow of data is the velocity vector.

4) Variety

We have noticed that photographs, sensor data, tweets, encrypted packets, and so on. Each of these is very different from each other. This data isn't the old rows and columns and database joins of our forefathers. It's very different from application to application, and much of it is unstructured. That way it doesn't easily fit into fields on a spreadsheet or a database application.

Take, for example, email messages. A legal discovery process might require sifting through thousands to millions of email messages in a collection. Not one of those messages is going to be exactly like another. Each one will consist of a sender's email address, a destination, plus a time stamp. All messages will have human-written text and possibly attachments.

Photos, videos and audio recordings, email messages, documents, books, presentations, tweets and ECG strips are all data, but unstructured, and varied incredibly.

All that data variety makes up the diversity vector of big data.

II. MANAGING THE THREE VS

It would take a library of books to explain all the various methods that big data practitioners use to process the three Vs. For now, though, our big takeaway should be this: once we start talking about data in terms that go beyond basic buckets, once you start talking about epic quantities, insane flow, and wide assortment, we're talking about big data.

One final thought: there are now ways to sift through all that insanity and glean insights that can be applied to solving problems, discerning patterns, and identifying opportunities. That process is called analytics, and it's why, when discussed data, we often hear the term analytics applied in the same sentence.

The three Vs explain the data to be analysed. Analytics is the process of deriving value from that data. Taken together, there is the potential for amazing insight or worrisome oversight. Like every other great power, big data comes with great promise and great responsibility.
III. TWO NEW EMERGING VS

Veracity defines quality and origin of data.
Value is the derived outcome from big data study and understanding.

Tools typically used in big data scenario
1. NoSQL
   - Databases: MongoDB
   - CouchDB
2. MapReduce
   - Hadoop
   - Hive
3. Storage
   - S3
   - Hadoop Distributed File System
4. Servers
   - EC2
   - Google App Engine
5. Processing
   - Yahoo! Pipes

IV. HADOOP FOR BIG DATA

What is Hadoop for big data?
Hadoop is the center platform for structuring Big Data, and solves the problem of formatting it for succeeding analytics purposes. Hadoop is an example of distributed computing architecture which consisting of multiple servers using commodity hardware, making it relatively inexpensive to scale and support extremely large data stores.

- Apache Hadoop and MapReduce on Amazon EMR.
- Hadoop Distributed File System vs. ...
- Data Types, Readers, Writers and Splitters.
- Data Mining and Filtering.
- Shell Comments and HDFS.
- Cloudera, Hortonworks and Apache Bigtop Virtual Machines.

Applications
Big data can be applied to
- Real-Time Fraud Detection
- Complex Competitive Analysis
- Call Center Optimization
- Consumer Sentiment Analysis
- Intelligent Traffic Management And
- To Manage Smart Power Grids

To name are only a few applications. In short, big data simply means more than an organization can manage effectively with their current BI program. But with the right analytics, big data can deliver richer insight since it draws from multiple sources and transactions to uncover hidden patterns and relationships.

There are four types of big data BI that really aid business:
1. Prescriptive – This type of investigation reveals what actions should be taken. This is the most valuable kind of investigation and usually results in rules and recommendations for next steps.
2. Predictive – An analysis of likely scenarios of what might happen.
3. Diagnostic – A look at past performance to determine what happened and why. The result of the analysis is often an analytic dashboard.
4. Descriptive – What is happening now based on incoming data. To excavate the analytics, you typically use a real-time dashboard and/or email reports.

V. THE FUTURE OF BIG DATA

No one declined concept of big data that it has taken the business world by storm, but question arise what’s next? Will data continue to grow? What technologies will develop around it? Or will big data in next trend cognitive technology become a relic as quickly? Or fast data?

Let’s look at some of the predictions from the foremost experts in the field, and how likely they are to come to pass.
1. Data volumes will continue to grow. In present day of internet world, we will continue generating bulk amount of data, so the number of devices handheld and Internet-connected devices exponentially grows.
2. Ways to analyze data will improve. As Ovum Says, While SQL as the standard, Spark is emerging as an analytical complementary tool and will continue to grow.
3. **More tools for analysis (without the analyst) will emerge.** Microsoft MSFT+0.18% and Sales force both recently announced features to let non-coders create apps to view business data.

4. **Prescriptive analytics will be built into business analytics software.** IDC predicts that before 2020 intelligence will include in half of all business analytics software. Users will want to be able to use data to make decisions in real time programs like Kafka and Spark.

5. **Big data will face huge challenges around privacy,** especially with the new privacy regulation by the European Union. Companies will be forced to address the ‘elephant in the room’ around their privacy controls and procedures. According to Gartner, business ethics violations will be related to data is about 50% by 2018.

6. **“Autonomous agents and things” will continue to be a huge trend,** according to Gartner, including robots, autonomous vehicles, virtual personal assistants, and smart advisers.

7. **Cognitive technology will be the new buzzword.**

8. **Big data for development** According to Forrester, more companies will attempt to drive value and revenue from their data.

9. **Big data will replace by “Fast data” and “actionable data”,** Experts says that, big not necessarily better when it comes to data, and that

**REFERENCES**


