

Report

by Josh Benjamin Rowland

FILE	2664_JOSH_BENJAMIN_ROWLAND_REPORT_93287_396469451.DOCX		
TIME SUBMITTED	22-MAR-2018 11:44 AM (UTC+0000)	WORD COUNT	1444
SUBMISSION ID	85300670	CHARACTER COUNT	7309

Josh Rowland
P145133
Emerging Web Technologies
Emerging Web Tech
Michael Dacey

Big Data

Big data is of massive importance in the current age where everything is online so companies both big and small need a good method of handling data and ensuring that the four v's are met. Big data can vary from place to place as one companies big data might be 10PB while another could be 10TB, there is no set amount to class as big data[1]. When analysing big data, we must consider the four V's:[1]

- Volume – What makes big data big, the sheer amount of storage space needed for everything so there is no point having a minimum storage amount as nobody is sure how much data is being generated and how much information is truly being collected, so you need enough storage to store everything.
- Velocity – This is the amount of incoming data that needs to be processed. There are countless actions that needs to be process every minute of every day such as Facebook messages or card payments made every minute. The company must have a way to process all of these actions.
- Variety – A goal of big data is take unstructured data like images and recording and make sense of it when comparing it to structured data such as bank statements. Structured data has rules set to it while unstructured data has no rules and can be anything.
- Veracity – Having good and trustworthy data is necessary when collecting any data, not just big data. If the data is wrong, then the analysis will lead to incorrect information and cause problems for the company.

There are many big data web applications online to use however we will look at the Amazon web services for big data or AWS. Rather than being a single application AWS is a marketplace full of software and frameworks to use for many things big data included, it allows you to create an application that revolves around the software or you can implement it into an already existing system.[2] AWS lets the user build any analytics application regardless of the volume, velocity and variety of the data and due to the broad amount of features it provides in order to analyse the data[2]. AWS provides features such as analytics frameworks like the Amazon Athena interactive query search which allows the user to analyse petabytes of data and as there are no warehouses it can be analysed immediately. Volume is one of the main issues when handling big data as you need a near unlimited amount of storage space which AWS provides through its many different types of storage options however because they have so many option it may take time to find what storage method such as the Amazon DynamoDB which has default limits but the user can pay more to increase their data limits to the exact amount that they require[2]. The velocity of the data is another major aspect of big data that needs to be handled because large companies like Amazon who process countless card payments every minute. They need a way to effectively analyse which is where the AWS Kinesis options can be used to analyse the velocity of the data like the Kinesis Firehose which allows the user to load massive amounts of streaming data in the AWS and enable near real-time analytics which takes the input information and prepares and loads it into your chosen data storage option which is then converted into easy to analyse data using analytics tools[3]. The variety of big data is a more complex issue as it is split into structured and unstructured data. The AWS S3 is a great way to store and analyse a variety of data as it can store any time of data from anywhere[4]. The veracity of the data is

different as it is difficult to control unlike the other three v's as the users are the main problem as they may enter data incorrectly either by mistake or on purpose which can lead to a bad veracity of data. The best way to prevent mistakes is to simply use error checking methods throughout the system from the moment the user inputs data like an email address to the output of when analysing data. The AWS Apache Spark on Amazon EMR is a popular software for dealing with it as it allows the user to filter the data in real-time and siphon the useful data and cast away the useless data, however it is the issue that is hardest to deal with when handling big data [5]. AWS is very useful as it has so many options to choose from to suit your exact needs while only paying for what you use while covering the four v's of big data, similar to Hadoop it is basically a large ecosystem that is incredibly flexible and dynamic.

Hadoop Hive

Before discussing Hive we must first mention Hadoop which is a software framework for distributed computing known commonly as HDFS or Hadoop Distributed File System and big data processing using MapReduce[6]. MapReduce as mentioned above is actually a programming model created by Google for data processing big data where the users specify a map function that has a key/value pair that generates a set of intermediate key/values pairs, while the reduce function is used to merge all the intermediate values that have the same intermediate key[7]. Hadoop splits files into large blocks of data and distributes them across separate nodes that are connected to the Hadoop system[8]. There are many add-ons to the Hadoop ecosystem such as Avro, Pig, Tez and more however we will discuss Hive[6].

Hive is a data warehouse project that is used on top of Hadoop that provides data analysis, queries and summarisations[9]. Hive provides an interface similar to SQL to allow the user to query data that may be stored in several different databases or systems that are connected to the Hadoop ecosystem[9].

Hive's primary use is for large data sets or big data and has features such as:[10]

- mechanisms to structure a variety of data formats in order to organise and make the data easier to understand.
- Several tools that enable easy access to data with SQL which enables data warehousing operations such as ETL or extract/transform/load and also reporting and data analysis.
- query execution using MapReduce, Apache Tez or Spark.
- Access to data stored in Apache HDFS or in other storage systems like Apache HBase.
- Hive also has a procedural language with HQL-SQL and sub-second query retrieval using Hive LLAP, Apache YARN or Slider.

Hive is incredibly useful as it allows the user to extract information from Hadoop clusters using SQL which is a very popular because it's an effective method of organising and using the data[9]. Hive does the heavy lifting of mapping operations to MapReduce which allows the user to focus on the queries which are then translated by Hive into MapReduce jobs thus exploiting the scalability of Hadoop while keeping it simple using SQL[9]. Hive is most

suited for data warehouse applications where a majority of static information is analysed such as big data that is maintained minded for information and insights, not where the data is rapidly changing [9]. Hive is also useful for developers as those who know SQL can learn Hive easily and port SQL based application over to Hadoop simpler than any other Hadoop language or tools[9]. To summarise Hive is ideal for any company that deals with big static data as long as the developers have some experience with SQL as it revolves around SQL and it can be stressful for developers to learn a new language in a rush which can lead to issues and errors.

Testing

Assignment

Sending message

Assignment

Read Last message

localhost:54738 says
Got message
status: success
requestStatus: 200

Assignment

Send Message

Read Last Message

Test

Read Message

Edit Message

Delete Message

Read All

Reading from an ID

Assignment

Send Message

Read Last Message

test1

Read Message

Edit Message

Delete Message

Read All

Editing

Assignment

test1

Modified

Assignment

This is test1

Modified

Deleting

Assignment

argr

1

localhost:54738 says

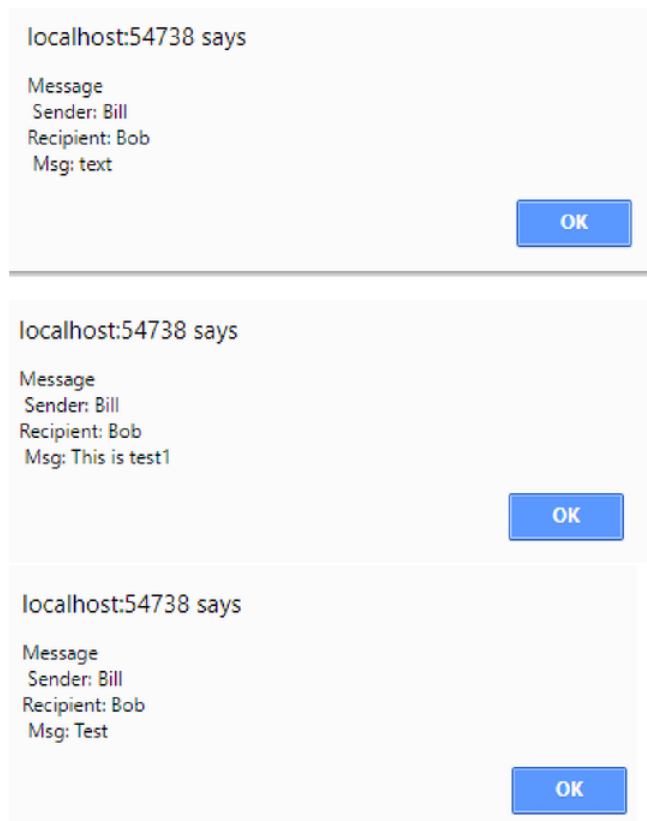
Deleted message
status: success
requestStatus: 200

Read all

Assignment

localhost:54738 says

Got message
status: success
requestStatus: 200 num of messages:3



References

- [1] J. Hurwitz, M. Kaufman, F. Halper, and D. Kirsch, *Big Data For Dummies*. Wiley, 2013.
- [2] "Big Data Solutions – Amazon Web Services (AWS)." [Online]. Available: <https://aws.amazon.com/big-data/>. [Accessed: 03-Mar-2018].
- [3] "Amazon Kinesis Data Firehose – load data streams to data stores." [Online]. Available: <https://aws.amazon.com/kinesis/data-firehose/>. [Accessed: 03-Mar-2018].
- [4] "Cloud Object Storage | Store & Retrieve Data Anywhere | Amazon Simple Storage Service." [Online]. Available: <https://aws.amazon.com/s3/>. [Accessed: 05-Mar-2018].
- [5] "What is Apache Spark? – Amazon Web Services (AWS)." [Online]. Available: <https://aws.amazon.com/emr/details/spark/>. [Accessed: 05-Mar-2018].
- [6] "Welcome to ApacheTM Hadoop®!" [Online]. Available: <http://hadoop.apache.org/>. [Accessed: 07-Mar-2018].

- [7] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters."
- [8] "What is HDFS? - Apache Hadoop® Distributed File System | IBM Analytics." [Online]. Available: <https://www.ibm.com/analytics/hadoop/hdfs>. [Accessed: 07-Mar-2018].
- [9] E. C. D. W. J. Rutherglen, *Programming Hive*. O'Reilly Media, Inc, 2012.
- [10] "Home - Apache Hive - Apache Software Foundation." [Online]. Available: <https://cwiki.apache.org/confluence/display/Hive/Home>. [Accessed: 06-Mar-2018].

Report

GRADEMARK REPORT

FINAL GRADE

40/100

GENERAL COMMENTS

Instructor

Big data application. (20) 19

Discussion of Hadoop technology (20) 18

Quality of website (Typescript) (20) 0

Single page application (20) 0

Dynamic content using AJAX (15) 0

Software testing (5) 3

Review of the four V's is a bit brief and doesn't evaluate the impact of these factors. However, you have provided some discussion on AWS. as it relates to the these factors which is quite good. Good evaluation of Hive.

Code is just a copy of that provided in class.

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9