ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Analytics on Command Centre Data in Healthcare Systems: A Case Study Implemented using Apache Hadoop, Avro and Crunch.

Archana M Kanthi[1], Annapurna P Patil[2]

M.Tech Student, Dept. of CSE, MSRIT, Bangalore, India[1]

Professor, Dept. of CSE, MSRIT, Bangalore, India[2]

ABSTRACT: In day today life, large amount of data is generated, this data can be either structured or unstructured. It is important to deal with the data, which is generated at a greater speed and quantity. Data Analytics helps to deal with such data; it examines raw data to develop a conclusion which will be used for making better business decisions.The demand for data analytics in health care sector is growing rapidly. The command centers used in many industries like military, space, government, broadcast entertainment, is also introduced in health care sector. In recent years, countries like U.S, U.K, Brazil, Canada and Australia has introduced the idea of command centers in health care industries. Health care solutions have acquired an organization that provides solutions for health care workforce management. Health care company has built a new management system for command center which includes six functionalities like patient flow, operations dashboard, real time location service, acuity, department management and scheduling. There is no real time data to test this system since it is a new application introduced for first time.The main objective of this project is to generate real time random data for schedule module used in command center using apache Avro and process this data to get the target data that is number of skilled person's required using apache Crunch. This would enable validation of command center ETL (Extract, Transform and Load), processing and dashboard applications.

KEYWORDS:Schedule module; Apache Maven; Apache Avro; Apache Crunch.

I. INTRODUCTION

Nowadays large amount of data is generated and used by the application and users. Data Analytics is a process of examining the raw data and processing according to the application and requirement imposed by the user. Data analytics mainly deals with raw data to provide a conclusion which will be helpful for business decisions in an organization. It is used in almost all fields like education sector, health care sector, business sector and many more. Data analytics is very important to analyse large amount of data. This project mainly concentrates on data analytics in health care sector. Initially in hospitals, record keeping and regularity requirements of medical and healthcare functions, like clinical decision support, disease surveillance, and population health management was done using hard copy [1]. But the current trend is digitizing all these records and requirements using data analytics. Data analytics in health care sector helps to reduce administrative cost and increase in improved outcome. One fourth of the health care budget is going to administrative section because of usage of man power [2-5]. It also reduces fraud, misuse of patient's information and also improves patient wellness because of better care co-ordination.

Big data in healthcare refers to electronic health data sets which are so large, complex and difficult (or impossible) to manage with traditional software and/or hardware. Big data in healthcare is increasing day by day not only because of its volume but also because of the variety of data and the speed at which it should be managed. Health data volume is expected to grow high in coming years. In addition, healthcare reimbursement models that are changing in a meaningful manner to use and pay for performance, emerging as critical new factors in today's healthcare environment. Although profit should not be a primary motivator, it is very important for healthcare organizations to use the available

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

tools, infrastructure, and techniques in an efficient manner or else there will be risk of losing millions of dollars in revenue and profits.

The hospital command center provides required resources in emergency and nonemergency condition like if any disaster happens instructions will be given after the incident, which includes evacuation details, emergency service details, and news updates. That time there is need of certain equipment's, nursing staff, and necessary supplies such as food and water this all will be provided by command center. It also deals with internal emergencies like power cut off, work place violence etc. Health care solution has acquired an organization that provides solutions for health care workforce management. Command center is a new management system that consists of six functionalities they are: patient flow, operations dashboard, real time location service, acuity, scheduling and departmental management. This paper mainly deals with scheduling part.

This paper involves case study of data generation and data processing using Apache Hadoop, Apache Avro and Apache Crunch in maven environment for schedule module used in command center. The detailed study of this will be explained in further sections and chapters. Firstly specification of area, problem definition, objective and its scope are explained in this section. Later, further chapters includes details about tools used in this paper, system design and architecture, system implementation actually how paper was implemented and finally the outcome.

II. LITERATURE SURVEY

Data in health care sector is growing in a faster manner, the present data analytical methods can be applied, but for the data of patient health which is not analyzed and it is needed to be understood for diagnosing purpose. If the data is analyzed effectively using big data tools then there is a chance of detecting diseases in earlier stage and cures it. Data analytics can answer many questions like patient can know which treatment can help him to stay for longer period, how much percent of complications, risk will be present in the chosen surgery. Big data of health care will define all 3v's in big data they are: volume, variety and velocity. Large volume of data in health care sector is accumulating day by day with high velocity and huge variety [6]. To analyze this data a conceptual architecture of big data analytics is defined, as shown in below Figure 1.

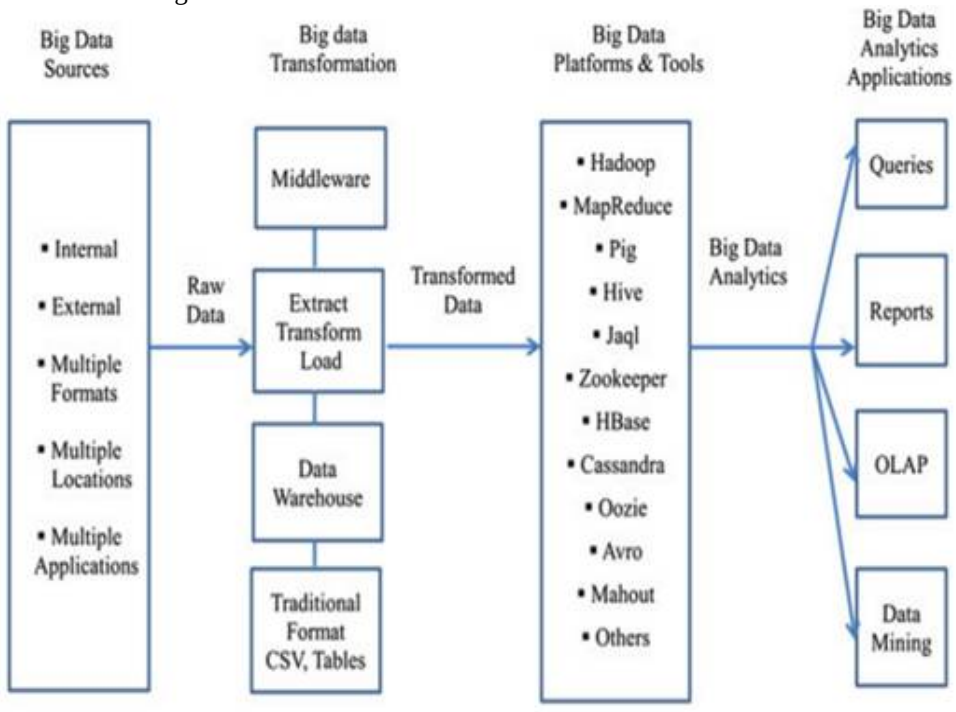

Figure 1: Conceptual architecture of big data analytics in healthcare sector [6].

Big data in health care sector can be obtained from many sources and in any format. Sources can be internal or external, the internal sources can be electronic health record and external sources can be government sources. Multiple

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

formats can be csv files, relational database, and text format. The raw data is first converted into transformed data and then analyzed using big data tools to get reports.

I Tools Used

A. Apache Maven

It is a project management tool. It is based on POM.xml file, which is project object model. It can manage project build process, reporting and documentation from central piece of information [7]. The result of this tool can be used building and managing java based project [8].

The main goal of Apache Maven is to provide a development environment for a developer to develop a project within a shortest period of time. To reach this goal there are certain features provided by the Apache Maven. They are as follows:

1. Makes the build process easy
2. Provides a uniform build system
3. Provides quality project information
4. Provides guidelines for best practices development
5. Allows transparent migration to new features

B. Apache Avro

Apache Avro is data serialization system [9]. Data serialization is a process in which data is converted into binary format or textual format; so that it can be stored in a buffer or memory of a computer or it can be transferred through the network. It provides rich data structures, compact and fast binary data format and container file to store persistent data, remote procedure call and integration with different languages. Even code generation is optional, it is not required to read and write files.

C. Apache Crunch

Apache Crunch is a high level tool for writing data pipelines. It develops java libraries, which helps developers to build pipelines and applications with better performance and testability. Apache Crunch is used to solve the problems which are working with non-relational data like complex records, HBase tables, vectors, geospatial data, etc. and which requires lots of custom logic to be written via user-defined functions.

The input data for Apache Crunch is taken from source folder and the output data is stored in a target folder. PCollection acts as a data container abstraction, data format and serialization is done using plain old java objects (POJO) and PTypes. Data transformation is done using (DoFn) do functions.

III. SYSTEM DESIGN

System design is the most important part to be considered before the start of the project. It describes architecture, modules and components used in the system and the different interfaces connected to the components along with the data.The Figure 2 shows the layered architecture of the project in which Apache Crunch is built on top of Hadoop.

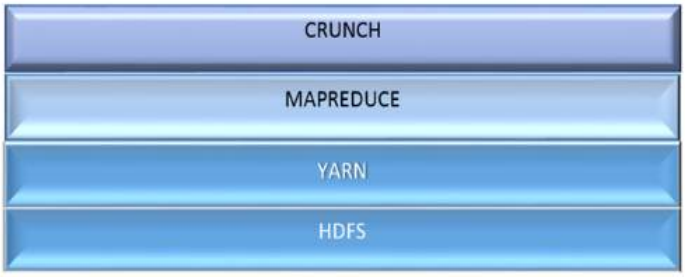

Figure 2: Layered architecture.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

In this architecture crunch pipeline is built on top of the Hadoop distributed file system. Apache Crunch helps developer to deal with complex logic and process the data faster. Hadoop 0.20.2 is installed and on top of that Crunch 0.9.0 is built using maven environment.

I Description of Schedule Module

Table 1: Schedule table.

schedule							
schedule id	Unique id	schedule type	cost center id	participant	effective period	source	active

The schedule module represents a schedule table as shown in Table 1. A schedule controls the date and time available for the actor. The schedule belongs to a single owner, the actor, which is normally a practitioner, location or device. It contains schedule id, unique id, schedule type, cost centre id, participant, effective period, source and active entity. The schedule id is a unique identifier of the schedule as provided by the data source. A unique id is the unique identifier for schedule across the whole reference record. The schedule type can be used for the categorization of healthcare services or other appointment types. The cost centre id is the cost centre associated with schedule as provided by the data source. The participant is the one who is associated with this schedule that is the nurse assigned to a cost centre. Effective period is the time period for when the schedule did (or is intended to) come into effect or end. Source is the source information for the schedule data. Active entity is an indication whether the entity is active or inactive as determine by the data source. If no indication was provided from the data source, the entity is assumed to be active.

Schedule module has sub modules like schedule type, participant, effective period and source. The schedule type is a record of type code which is an aggregation of a raw and a mapped standard code. Participant is a record of type participant which is a structure that encapsulates different types of appointment participants, such as provider or resource (e.g. room). Effective period is a record of type period which contains start date and end date. Source is a record of type source, which is structured representing information about the concept as it pertains to the source system, which it is originated from.

II Detailed Design

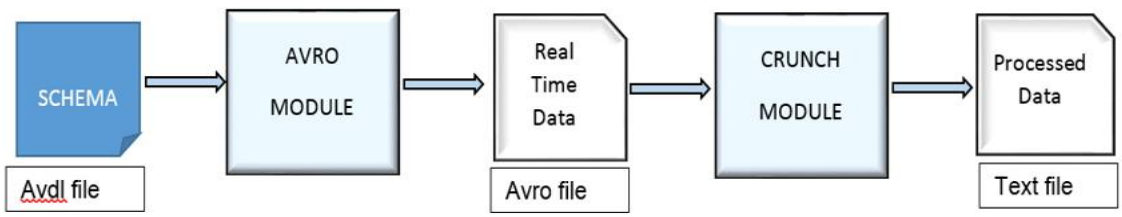

Figure 3: Detailed design.

The figure 3 shows the detailed design of the system. The project contains two main modules one is Avro module and another is Crunch module. Avro module is used for generating data for schedule module. Crunch module is used for processing the generated data to get the target data.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Schemas are given as input to Avro module, here schemas are written in AVDL format. This schema contains schema for schedule module. Depending on schema the classes are generated and a new class is created to generate a real time data for schedule module. In this new class variables are set using random data utils, data for schedule module and later serialize and de-serialize it. Finally the generated data are stored in Avro file.

The generated Avro file acts as input to the Crunch module. Crunch module processes the generated data using mem pipeline and logic written inside the do function, which generates target data that displays number of skilled persons.

IV. SYSTEM IMPLEMENTATION

System implementation explains the exact implementation of the application that is how the application was built from initial stage to final stage.

The steps involved in the system implementation are mentioned as follows:

Step 1: Setting up a Maven environment in Eclipse.
Step 2: Adding required dependencies in POM.xml file.
Step 3: Writing AVDL schemas for schedule module.
Step 4: Generating random data for schedule module and serializing and deserializing it using Apache Avro.
Step 5: Processing the generated data for schedule module to get required number of skilled person in hospitals.

V. RESULTS

The outcome of the implemented system is to generate a random data for schedule module using Apache Avro and processing it to get a target data using Apache Crunch. Schedule module defines the attributes required for managing daily staff/skill allocation in a particular facility for a specified period. It consists of schedule id, uid, schedule type, cost center id, participant details, effective period, source data, and active entity. After the generation of data for schedule module, the generated data is processed to get target data.

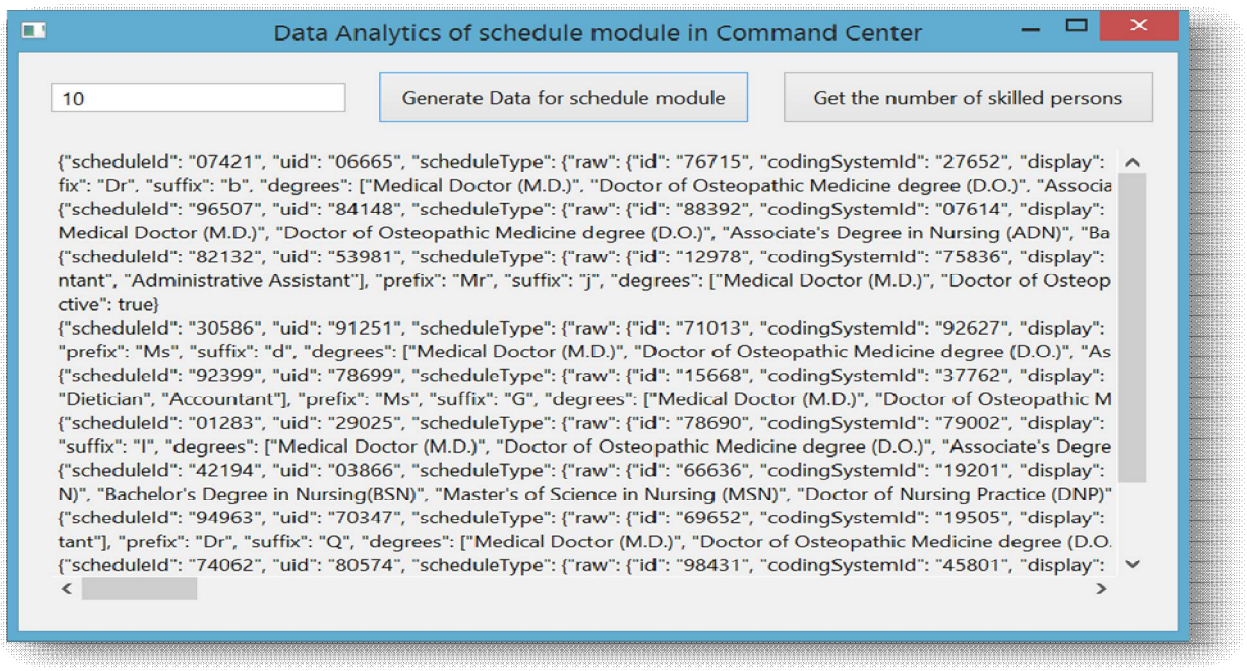

Figure 4: Avro generated data for schedule module.

The number of records for schedule module to be generated should be entered into the text box. For example if 1000 records are needed to be generated then enter 1000 in the text box. To get the Avro generated data for schedule model

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

button1 is clicked that is “Generate data for Schedule Module” and to get the Crunch processed data for number of skilled persons, second button is clicked that is “Get the number of skilled persons”.

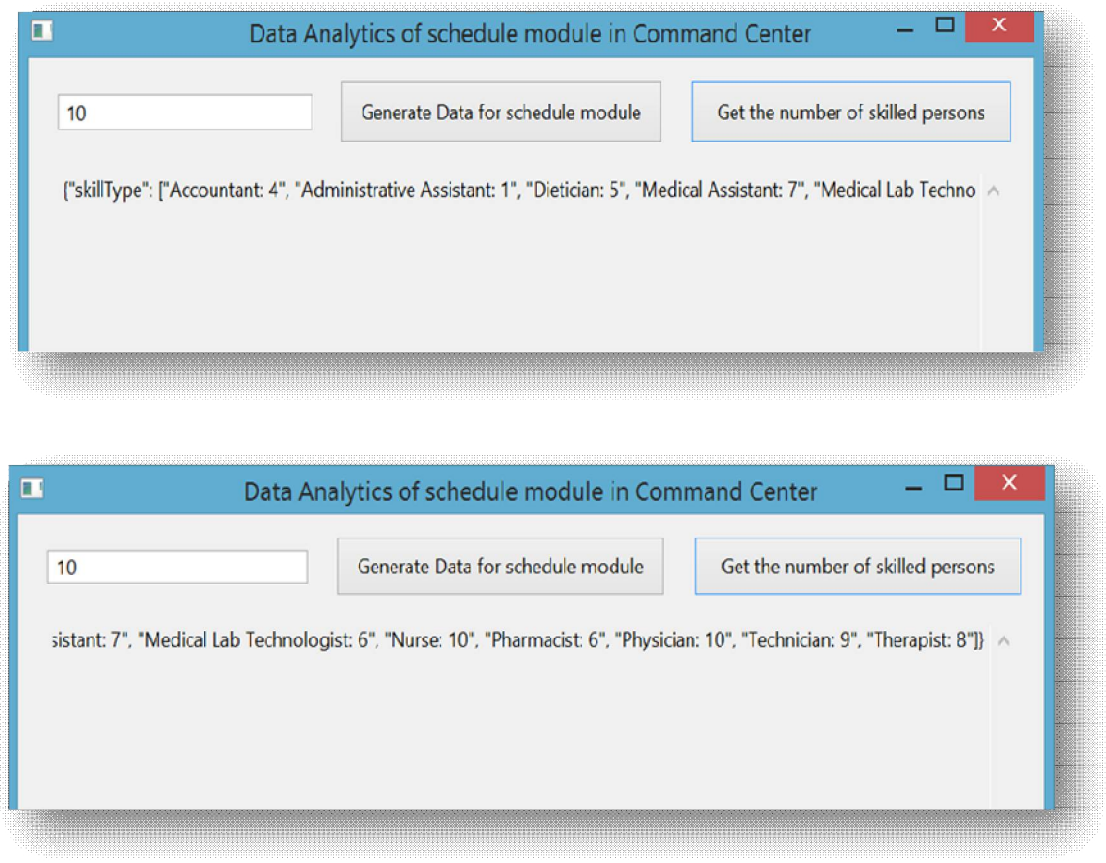

Figure 5: Crunch processed data to get the number of skilled persons.

The following Figure 4 and 5 shows the Avro generated data and Crunch processed data for schedule module in GUI. Figure 4 shows the generation of 10 records of schedule module. Figure 5 shows the total number of skilled persons required for generated 10 records of schedule module.

Table 2: Time taken to generate different number of records for schedule module.

Data Sets Generated	Time Taken
1,000 records	10 sec
10,000 records	20sec
1,00,000 records	4min
10,00,000 records	34min

The table 2 shows the time taken to generate data for schedule module, that is Avro module took 10 seconds to generate 1,000 records of schedule module, 20 seconds for 10,000 records, 4 minutes for 1,00,000 records and 34 minutes for 10,00,000 records

Crunch pipeline takes few seconds to process 1, 00,000 of data. Thus using Apache Avro, Apache Crunch and Apache Hadoop has made data generation and data processing fast. Apache Maven has made the build process of the project easy. Thus the technology has brought a new approach in a data analytics.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

VI. CONCLUSION AND FUTURE WORK

Hospital command center is a new concept introduced recently in Health care sector. Health care solution has acquired an organization that provides solutions for health care workforce management. Command center is a new management system that contains six functionalities like patient flow, operation dashboard, real time location service, acuity, and scheduling and departure management. There is no data to test this new system. Analytics on command center using apache Hadoop, apache Avro and apache Crunch for operational excellence of command center has solved the existing problem.

The main area considered in this project is staff utilization. This will allow enterprise visibility enabling predictions to help and forecast staff needs, acuity impacts and engage staffing office in resource management. This reflects high cost resources and over/under staffing.

The outcome of the implemented system is the generation of random data for schedule model that defines the attributes required for managing daily staff/skill allocation in a particular facility for a specified period using apache Avro. The generated data is processed using apache Crunch pipeline to get the number of skilled persons required at that moment. The big data tools used in this project for data analytics improves the efficiency and speed of generating and analyzing of data.

The technologies used in this project, can be used in command centers present in military, government and broadcast entertainment, so that there will be improvement in the co-ordination between the teams through a command center.

REFERENCES

1. Raghupathi W,"Data Mining in Health Care. In Healthcare Informatics", Improving Efficiency and Productivity, 2010:211-213.
2. Burghard C, "Big Data and Analytics Key to Accountable Care Success", IDC Health Insights, 2012.
3. Dembosky A, "Data Prescription for Better Healthcare", Financial Times, December 12, 2012.
4. Feldman B, Martin EM, Skotnes T,"Big Data in Healthcare Hype and Hope".
5. Lorraine Fernandes, Michele O'Connor and Victoria Weaver, "Big Data, Bigger Outcomes", Journal of AHIMA 83, no.10 (October 2012): 38-43.
6. Wullianallur Raghupathi author and Viju Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems,DOI: 10.1186/2047-2501-2-3, 5 January 2014.
7. Apache Maven Project, [Online] Available: https://maven.apache.org/what-is-maven.html
8. Ashay Patil, Maven Architecture: How does Maven Works?, Aug 2, 2015.
9. Apache Avro ™ 1.8.1 Documentation, [Online] Available: https://avro.apache.org/docs/current/
10. Apache Avro ™ 1.8.1 Documentation, [Online] Available: https://avro.apache.org/docs/current/gettingstartedjava.html
11. Data serialization system, [Online] Available:http://www.tutorialspoint.com/avro/avro_tutorial.pdf
12. Matt Pouttu-Clarke, "Cascading.Avro",January 13, 2011.
13. Micah Whitacre,Scaling People With Apache Crunch, May 9th, 2014.

BIOGRAPHY

Archana Mallikarjun Kanthi is a MTECH Student in CSE, M S Ramaiah Institute of Technology, Bangalore, Karnataka, India

Dr. Annapurna P Patil is working as Professor at Department of CSE, M S Ramaiah Institute of Technology, Bangalore, Karnataka, India