REPORT FILE

Contents

Chapter1: Reflection

1.1 Introduction

In this project, a big data is managed and big data means that when the data do not exist in one file. A company has the data of employees in many that could not be manageable easily. Unfortunately, their business has multiple areas which all have customer data specific to that area, and this is fragmented within the organization. For example, Credit Card data is only stored by the financial systems, employment within HR, etc. There is not a single cohesive record representing customers. The SMB is looking to unify these ahead of further data investigation, and to pool all this data together into a central database. The data provided for this assessment is mock data representing a typical customer-facing business, these involve data such as names, banking credentials, family attributes, etc. These data files are provided as a mixed modality in a variety of formats (CSV, JSON, XML, and TXT. The work herein requires the processing of these data into a homogenous record, aligning the same customers from different sources together, which are then automatically entered into a Relational Database System using modern tools & libraries.

1.2 Challenges and Solutions

It is very difficult to extract the data from many files when the data in each file is not in a format means some missing values and some other problems. Major problem is to extract the data from a file and to store it in object. Because when you extract the data from a file, it is in a form of a block. It is big challenge to store it in the parameters of class.
This challenge is sorted out using the method **spilt**. It splits the content of black that are read by a file into a list. This list can be further splits to get the desired value. Split method is used many times according to the requirements.

Chapter 2: Relational Database

The data of company is managed manually and using files processing, where the data is not efficient, inaccurate and inconsistence. The data is stored in many files. In each file, data is not in a specific format. Data is not organized in these files. We have to make a relational database in which data will be stored efficient, accurate and consistence.

As the data is in many files, each file contains Customer information and other different information.

CUSTOMER Table

FirstName	**SecondName**	**Age**	**Sex**	**Vehicle_ID**	**IBAN_Id**	**Job_ID**	**Message_Id**
Oliver	Brady	68	Male	V101	null	null	null
Denis	Jackson	35	Male	V102	null	null	null
Jannet	Whitaker	79	Male	null	GB108101	null	null

In CSV file, Customer and its vehicle data is stored. We have to make a relational database, in which foreign keys will come.

Vehicle Table

V_ID	V_MAKE	V_Model	V_Year	V_Type
V101	Mitsubishi	WRX	2003	Sedan
V102	Toyota	Canyon Regular	2011	Convertible
V103	Honda	1500 Crew Cab	1999	Sedan

In JSON file, data about the bank will store. And V_ID will be foreign key in Customer Table.

Bank_Account Table

IBAN_ID	Credid_Card_No	Adress City	Card_Start_Date	CARD_End_Date
GB108101	10292560484	Landon	01-08-2015	01-08-2020
GC90102	94920572002	UK	15-02-2013	15-02-2019

In XML file, data about the bank will store. And Job_ID will be foreign key in Customer Table.

JOB TABLE

JOD_ID	Retired	Salary
J101	TRUE	10923
J102	FALSE	20603
J103	TRUE	59839

In TXT file, data about the bank will store. And MESSAGE_ID will be foreign key in Customer Table.

Message_ID	MESSAGE
M101	Debra Wood phoned up at the weekend
M102	Howard Johnson Congratulations on the promotion

Chapter 3: Potential Big Data issues

When you solve big data problems, there are many problems that a programmer faces. As in the given files**(User_Data.csv, User_data.json, User_data.xml and User_data.csv),** there are many problems when we try to fetch the records from a file. This is because of that records are not in a specific format. In one record, there are many fields are empty and some errors. Somewhere, there is no **firstName** and **secondName** are not mentioned in the files. For these records, there is meaning of that records(instance).

The major big data problems are explained below.

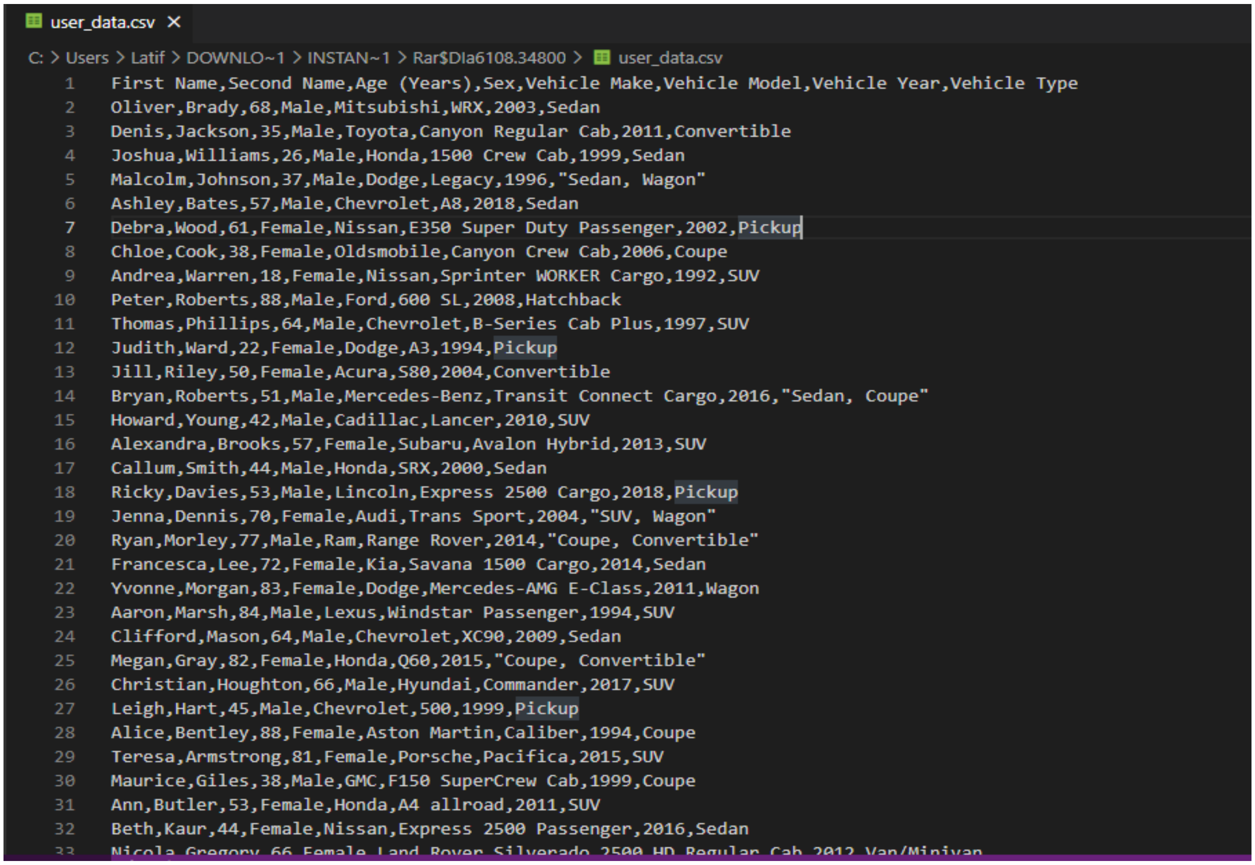
3.1 Understanding Lack

Organizations fail of their large information(Big_data) tasks because of insufficient understanding. Personnel might not recognize what facts is, its garage, processing, significance, and resources. Records experts can also know what is going on, but others may not have a clear photo. For example, if employees do no longer apprehend the importance of information garage, they might not maintain the backup of sensitive statistics. They won't use databases well for garage. As a result, whilst this vital facts is needed, it cannot be retrieved easily. To run those cutting-edge technology and massive information gear, organizations want professional records experts. these experts will encompass records scientists, information analysts, and statistics engineers to work with the gear and make feel of large data sets. One of the huge statistics challenges that any organisation face is a drag of lack of huge information experts. This is frequently because statistics managing equipment have developed unexpectedly, but in maximum cases, the specialists have not. Actionable steps were given to be taken to bridge this hole.

Companies can leverage information to enhance overall performance in lots of areas. some of the exceptional use cases for facts are to: lower fees, create innovation, release new merchandise, grow the bottom line, and boom performance, to call some. in spite of the benefits, agencies have been gradual to adopt facts generation or positioned a plan in vicinity for how to create a records-centric subculture.

One way to fight the slow adoption is to take a pinnacle-down method for introducing and schooling your company on information utilization and processes. in case your in-house crew doesn’t have the sources to take this on, consider bringing in IT professionals or specialists and conserving workshops to teach your organization.

SCREENSHOTS



user_data.csv
C: > Users > Latif > DOWNLO~1 > INSTAN~1 > Rar$DIa6108.34800 > user_data.csv

```
First Name,Second Name,Age (Years),Sex,Vehicle Make,Vehicle Model,Vehicle Year,Vehicle Type
Oliver,Brady,68,Male,Mitsubishi,WRX,2003,Sedan
Denis,Jackson,35,Male,Toyota,Canyon Regular Cab,2011,Convertible
Joshua,Williams,26,Male,Honda,1500 Crew Cab,1999,Sedan
Malcolm,Johnson,37,Male,Dodge,Legacy,1996,"Sedan, Wagon"
Ashley,Bates,57,Male,Chevrolet,A8,2018,Sedan
Debra,Wood,61,Female,Nissan,E350 Super Duty Passenger,2002,Pickup
Chloe,Cook,38,Female,Oldsmobile,Canyon Crew Cab,2006,Coupe
Andrea,Warren,18,Female,Nissan,Sprinter WORKER Cargo,1992,SUV
Peter,Roberts,88,Male,Ford,600 SL,2008,Hatchback
Thomas,Phillips,64,Male,Chevrolet,B-Series Cab Plus,1997,SUV
Judith,Ward,22,Female,Dodge,A3,1994,Pickup
Jill,Riley,50,Female,Acura,S80,2004,Convertible
Bryan,Roberts,51,Male,Mercedes-Benz,Transit Connect Cargo,2016,"Sedan, Coupe"
Howard,Young,42,Male,Cadillac,Lancer,2010,SUV
Alexandra,Brooks,57,Female,Subaru,Avalon Hybrid,2013,SUV
Callum,Smith,44,Male,Honda,SRX,2000,Sedan
Ricky,Davies,53,Male,Lincoln,Express 2500 Cargo,2018,Pickup
Jenna,Dennis,70,Female,Audi,Trans Sport,2004,"SUV, Wagon"
Ryan,Morley,77,Male,Ram,Range Rover,2014,"Coupe, Convertible"
Francesca,Lee,72,Female,Kia,Savana 1500 Cargo,2014,Sedan
Yvonne,Morgan,83,Female,Dodge,Mercedes-AMG E-Class,2011,Wagon
Aaron,Marsh,84,Male,Lexus,Windstar Passenger,1994,SUV
Clifford,Mason,64,Male,Chevrolet,XC90,2009,Sedan
Megan,Gray,82,Female,Honda,Q60,2015,"Coupe, Convertible"
Christian,Houghton,66,Male,Hyundai,Commander,2017,SUV
Leigh,Hart,45,Male,Chevrolet,500,1999,Pickup
Alice,Bentley,88,Female,Aston Martin,Caliber,1994,Coupe
Teresa,Armstrong,81,Female,Porsche,Pacifica,2015,SUV
Maurice,Giles,38,Male,GMC,F150 SuperCrew Cab,1999,Coupe
Ann,Butler,53,Female,Honda,A4 allroad,2011,SUV
Beth,Kaur,44,Female,Nissan,Express 2500 Passenger,2016,Sedan
Nicola,Gregory,66,Female,Land Rover,Silverado 2500 HD Regular Cab,2012,Van/Minivan
```

Picture 3.1.1

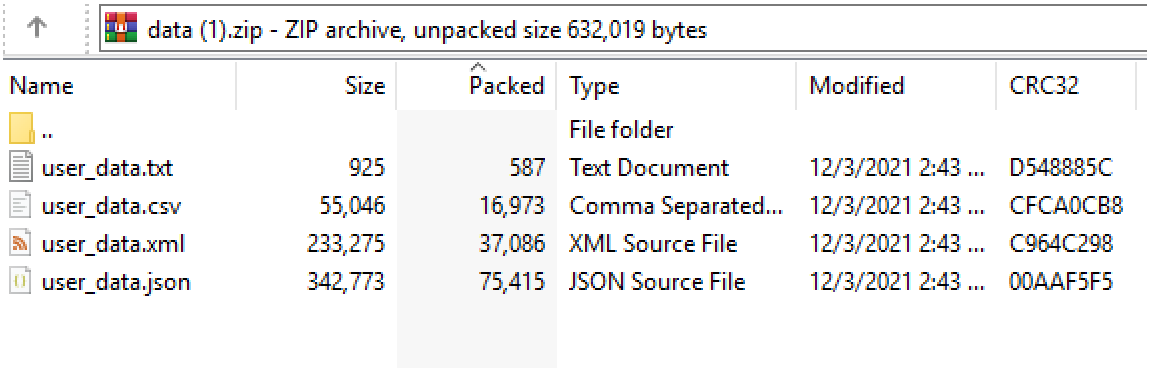
3.2 Growth Issue

One of the most pressing demanding situations of massive statistics is storing a majority of these large units of information nicely. The amount of facts being stored in information facilities and databases of corporations is growing hastily. As these information units grow exponentially with time, it receives extremely hard to address. Maximum of the information is unstructured and springs from documents files and other sources, which means that you can't find them in databases.

Businesses select current techniques to deal with these massive statistics sets, like compression, tiering, and deduplication. Compression is hired for reducing the range of bits inside the information, for that reason reducing its typical length. Deduplication is the system of getting rid of reproduction and undesirable information from a know-how set. Records tiering lets in groups to keep records in numerous storage levels. It guarantees that the information is residing inside the most appropriate space for storing. Statistics degrees are often public cloud, private cloud, and flash storage, relying on the data length and importance. Corporations are also deciding on big statistics equipment, like Hadoop, NoSQL, and other technologies.

In files, sometimes there is no more than one attribute value is mentioned. So, it is quite difficult to manage the files when the records in many files are no in a specific format.

SCREENSHOT



data (1).zip - ZIP archive, unpacked size 632,019 bytes

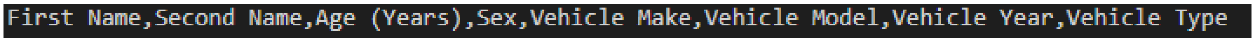
Name	Size	Packed	Type	Modified	CRC32
..			File folder		
user_data.txt	925	587	Text Document	12/3/2021 2:43 ...	D548885C
user_data.csv	55,046	16,973	Comma Separated...	12/3/2021 2:43 ...	CFCA0CB8
user_data.xml	233,275	37,086	XML Source File	12/3/2021 2:43 ...	C964C298
user_data.json	342,773	75,415	JSON Source File	12/3/2021 2:43 ...	00AAF5F5

Picture 3.2.1

3.3 Data Integrating from Many Resources

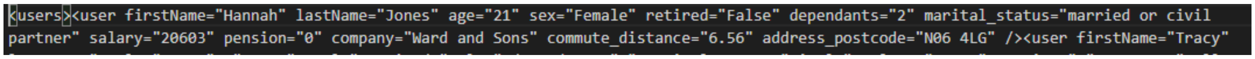
Records in an organization come from a selection of resources, consisting of many variables when there are so many to get the records. Combining all this information to put together reviews is a difficult project. This is a place frequently left out through firms. But, records integration is critical for analysis, reporting and business intelligence, so it needs to be perfect.

SCREENSHOTS



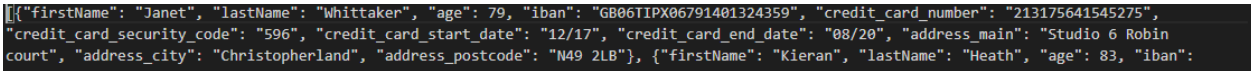
```
First Name,Second Name,Age (Years),Sex,Vehicle Make,Vehicle Model,Vehicle Year,Vehicle Type
```

Picture 3.3.1



```
<users><user firstName="Hannah" lastName="Jones" age="21" sex="Female" retired="False" dependants="2" marital_status="married or civil
partner" salary="20603" pension="0" company="Ward and Sons" commute_distance="6.56" address_postcode="N06 4LG" /><user firstName="Tracy"
```

Picture 3.3.2



```
[{"firstName": "Janet", "lastName": "Whittaker", "age": 79, "iban": "GB06TIPX06791401324359", "credit_card_number": "213175641545275",
"credit_card_security_code": "596", "credit_card_start_date": "12/17", "credit_card_end_date": "08/20", "address_main": "Studio 6 Robin
court", "address_city": "Christopherland", "address_postcode": "N49 2LB"}, {"firstName": "Kieran", "lastName": "Heath", "age": 83, "iban":
```

Picture 3.3.3

It is difficult to manage the manage the records different,

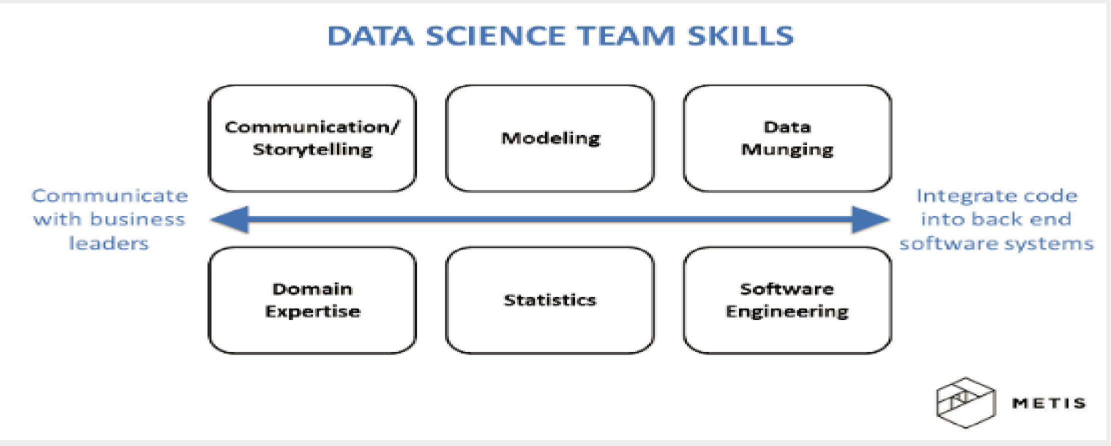
- Types of data resources
- Types of Data
- Data Format

3.4 Many Choices To Handle

Consistent with psychologist Barry Schwartz, much less actually may be greater. Coined because the "paradox of choice," Schwartz explains how option overload can motive state of no activity on behalf of a buyer. Alternatively, by proscribing a consumer's choices, anxiety and pressure can be lessened. Inside the world of information and data gear, the alternatives are almost as significant because the data itself. So it is understandably overwhelming while identifying the solution that's right for your commercial enterprise. Specially whilst it'll probably affect all departments and hopefully be an extended-time period strategy.

3.5 Professional Shortage

To run those contemporary technology and large information tools, corporations need skilled statistics experts. Those specialists will consist of information scientists, statistics analysts and data engineers who are skilled in working with the gear and making sense out of huge information units.


DATA SCIENCE TEAM SKILLS
Communication/ Storytelling
Modeling
Data Munging
Communicate with business leaders
Integrate code into back end software systems
Domain Expertise
Statistics
Software Engineering
METIS

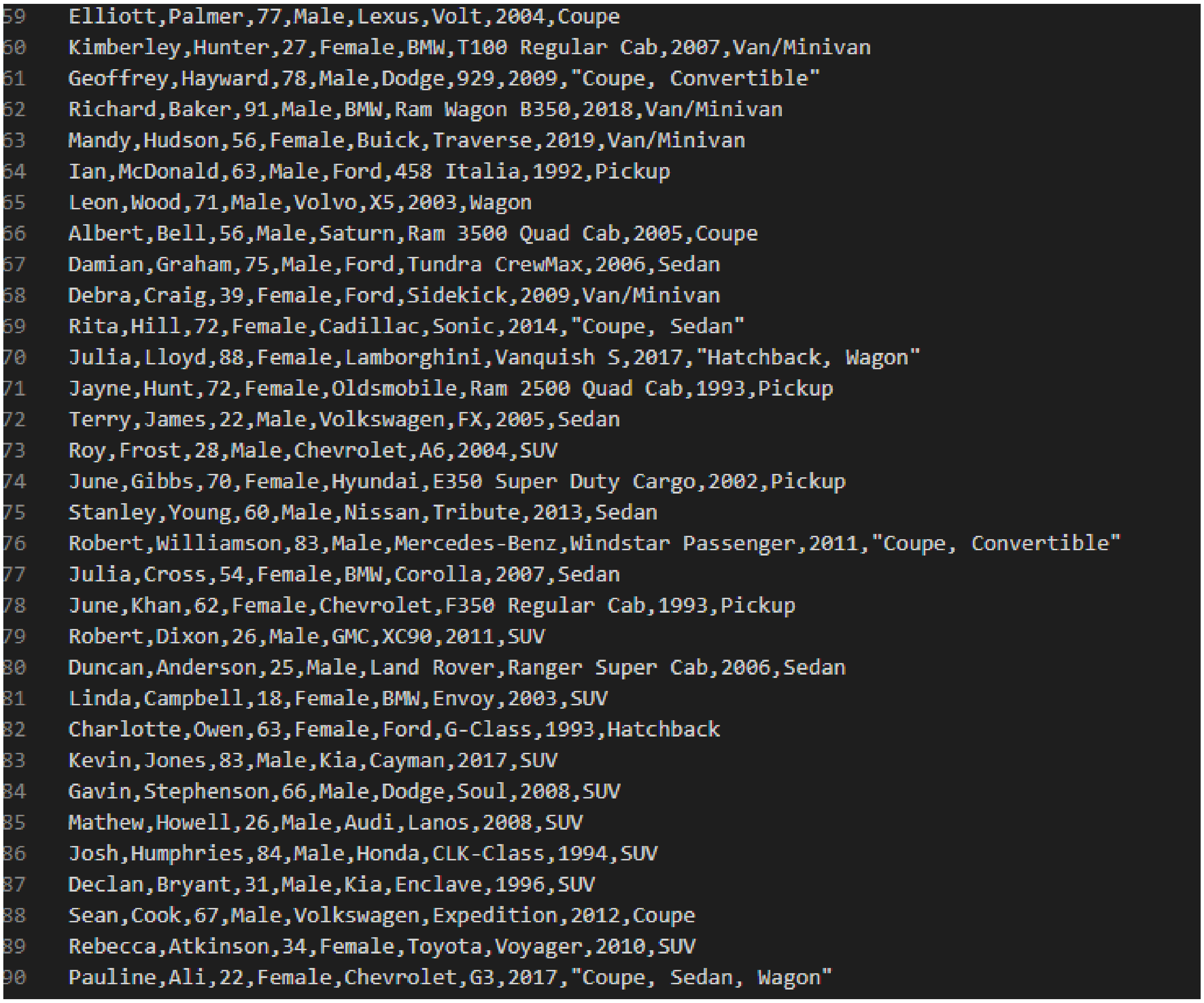
Picture 3.5.1

If the answer doesn't exist obviously, try and create it. while you may't manage how many records scientists and information analysts graduate every 12 months, you could leverage your modern body of workers and train the abilities you want them to have. You could additionally look for greater effective statistics equipment that make the analysis paintings much less complex, which open up recruitment to a broader pool of less specialized analysts.

3.6 Data Constantly Changing

Imposing the infrastructure and control of information can't be a fixed-and-forget venture. The character of facts is that it's constantly converting. Your client details and orders are always changing, as well as their interactions together with your enterprise.

Comprise records systems with superior device gaining knowledge of and interoperability a good way to adapt to the continuously converting panorama of statistics inputs, and in turn, outputs.



```
Elliott,Palmer,77,Male,Lexus,Volt,2004,Coupe
Kimberley,Hunter,27,Female,BMW,T100 Regular Cab,2007,Van/Minivan
Geoffrey,Hayward,78,Male,Dodge,929,2009,"Coupe, Convertible"
Richard,Baker,91,Male,BMW,Ram Wagon B350,2018,Van/Minivan
Mandy,Hudson,56,Female,Buick,Traverse,2019,Van/Minivan
Ian,McDonald,63,Male,Ford,458 Italia,1992,Pickup
Leon,Wood,71,Male,Volvo,X5,2003,Wagon
Albert,Bell,56,Male,Saturn,Ram 3500 Quad Cab,2005,Coupe
Damian,Graham,75,Male,Ford,Tundra CrewMax,2006,Sedan
Debra,Craig,39,Female,Ford,Sidekick,2009,Van/Minivan
Rita,Hill,72,Female,Cadillac,Sonic,2014,"Coupe, Sedan"
Julia,Lloyd,88,Female,Lamborghini,Vanquish S,2017,"Hatchback, Wagon"
Jayne,Hunt,72,Female,Oldsmobile,Ram 2500 Quad Cab,1993,Pickup
Terry,James,22,Male,Volkswagen,FX,2005,Sedan
Roy,Frost,28,Male,Chevrolet,A6,2004,SUV
June,Gibbs,70,Female,Hyundai,E350 Super Duty Cargo,2002,Pickup
Stanley,Young,60,Male,Nissan,Tribute,2013,Sedan
Robert,Williamson,83,Male,Mercedes-Benz,Windstar Passenger,2011,"Coupe, Convertible"
Julia,Cross,54,Female,BMW,Corolla,2007,Sedan
June,Khan,62,Female,Chevrolet,F350 Regular Cab,1993,Pickup
Robert,Dixon,26,Male,GMC,XC90,2011,SUV
Duncan,Anderson,25,Male,Land Rover,Ranger Super Cab,2006,Sedan
Linda,Campbell,18,Female,BMW,Envoy,2003,SUV
Charlotte,Owen,63,Female,Ford,G-Class,1993,Hatchback
Kevin,Jones,83,Male,Kia,Cayman,2017,SUV
Gavin,Stephenson,66,Male,Dodge,Soul,2008,SUV
Mathew,Howell,26,Male,Audi,Lanos,2008,SUV
Josh,Humphries,84,Male,Honda,CLK-Class,1994,SUV
Declan,Bryant,31,Male,Kia,Enclave,1996,SUV
Sean,Cook,67,Male,Volkswagen,Expedition,2012,Coupe
Rebecca,Atkinson,34,Female,Toyota,Voyager,2010,SUV
Pauline,Ali,22,Female,Chevrolet,G3,2017,"Coupe, Sedan, Wagon"
```

Picture 3.6.1

3.7 Securing Data

Agencies are recruiting more cybersecurity experts to guard their information. other steps taken for Securing big facts include: information encryption statistics segregation identification and get admission to control Implementation of endpoint protection actual-time safety tracking Use massive records protection tools, like IBM mum or dad. Securing those massive units of information is one of the daunting challenges of massive records. Frequently agencies are so busy in knowledge, storing and analyzing their information units that they push statistics protection for later levels. But, this isn't always a clever pass as unprotected facts repositories can emerge as breeding grounds for malicious hackers.

3.8 Managing Unstructured Data

Information control refers to the procedure of shooting, storing, organizing, and maintaining records gathered from diverse information units. The statistics units can be either dependent or unstructured and are available from a extensive range of sources that may include tweets, customer opinions, and internet of factors (IoT) information. Unstructured records presents an possibility to accumulate wealthy insights that could create a entire image of your clients and provide context for why sales are down or prices are going up.

The hassle is, coping with unstructured records at excessive volumes and high speeds approach which you're gathering loads of exquisite information but also a variety of noise that can difficult to understand the insights that add the maximum cost in your organisation. You may get ahead of huge facts troubles by addressing the following:

- What data needs to be integrated?
- what number of information silos need to be linked?
- What statistics are you hoping to beyou may get ahead of huge facts troubles by

3.9 Paying Loads of Money

Big facts adoption initiatives entail masses of fees. If you opt for an on-premises solution, you'll have to mind the fees of recent hardware, new hires (administrators and builders), energy and so on. Plus: despite the fact that the wanted frameworks are open-source, you'll nonetheless want to pay for the development, setup, configuration and maintenance of latest software program. If making a decision on a cloud-based huge facts solution, you'll nonetheless need to lease personnel (as above) and pay for cloud offerings, big statistics solution development in addition to setup and preservation of wanted frameworks.

Moreover, in both cases, you'll want to allow for destiny expansions to keep away from large records boom getting out of hand and costing you a fortune.

3.10 Compliance Hurdles

Whilst accumulating statistics, security and authorities rules come into play. With the rather recent introduction of the general statistics protection law (GDPR), it's even more important to apprehend the essential necessities for information series and safety, in addition to the results of failing to stick. Agencies need to be compliant and careful in how they use statistics to section clients as an example identifying which customer to prioritise or recognition on. Because of this the information need to: be a consultant pattern of consumers, algorithms need to prioritise equity, there's an knowledge of inherent bias in records, and massive statistics outcomes ought to be checked against historically implemented statistical practices.

3.11 Using Data for Meaning

You could have the information. It’s smooth, correct and organized. However, how do you use it to provide precious insights to improve your business? Many organizations are turning to sturdy information evaluation tools that could assist investigate the huge image, as well as damage down the statistics into meaningful bits of records which could then be converted into actionable effects.

Whether this means having a consistent reporting structure or a dedicated analytics team, be sure to turn your data into measurable outcomes. This means taking data and transforming into actions for the business to take in an effort to produce wins for the company.

References:

Kaisler, Stephen, et al. "Big data: Issues and challenges moving forward." *2013 46th Hawaii international conference on system sciences*. IEEE, 2013.

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January). Big data: Issues and challenges moving forward. In *2013 46th Hawaii international conference on system sciences* (pp. 995-1004). IEEE.

Katal, Avita, Mohammad Wazid, and Rayan H. Goudar. "Big data: issues, challenges, tools and good practices." *2013 Sixth international conference on contemporary computing (IC3)*. IEEE, 2013.

Chickerur, Satyadhyan, Anoop Goudar, and Ankita Kinnerkar. "Comparison of relational database with document-oriented database (mongodb) for big data applications." *2015 8th International Conference on Advanced Software Engineering & Its Applications (ASEA)*. IEEE, 2015.