REVIEW THE CHALLENGES OF FIVE HIGH LEVEL TASKS IN ENTERPRISE DATA ANALYSIS

[1]Sathish G, [2]Neethu G

[1]Assistant Professor, [2]II MCA Student
[1 2]Department of Computer Applications,
[1 2]Priyadarshini Engineering College, Vaniyambadi, Vellore, Tamilnadu, India

Abstract: This paper is about five high level enterprises of data analysis process discovery, wrangle, profile, model, report. Find that discovery and wrangle often the most tedious and time consuming aspects of an analysis as underserved by existing visualization and analysis tools these challenges are typically more acute with in large organizations with a diverse and distributed set of data sources. Finally this paper discusses about high level task and challenges in data analysis process.

Index Terms- Data, Analysis,Visualization.

I. Introduction

Organizations gather increasingly large and complex data sets each year. Analysts often work as a part of an analysis team or business unit. In this paper is about five high level enterprise of data analysis process.

Discover
The analysis acquired data necessary to complete their task within a large organization.

Wrangle
The analyst discovered the appropriate data to use it often need to manipulate the acquired data before it could use it for the analysis.

Profile
The analyst enters the phase of diagnosing the data quality issue & understands what they can make their data**.**

Model
After all the required data was assembled &understood, analyst could begin modeling their data**.**

Report
Analyst typically reported insight gained from modeling to the analyst of a business process (or) business unit.

These challenges are typically more acute within the large organizations with a divers and distributed set of data sources. We conclude the characterized the tasks within the analysis process of the work. Not all analysis requires all five tasks, and not all analysis performs each of them.

2. RELATED WORK

Many researchers have studied analysts and their processes within intelligence agencies. Although there is much overlap in the high-level analytic process of intelligence and enterprise analysts, these analysts often work on different types of data with different analytic goals and therefore perform different low-level tasks. For example, enterprise analysts more often perform analysis on structured data than on document and emails. In previous process is the scale and diversity of data source increase within enterprises, there is an opportunity for visual analytic tools to improve the quality of analysis and the speed at which it takes place. Tools that simplify tasks across the analytic pipeline could empower nonprogrammers to apply their statistical training and domain expertise to large, diverse data sets. Tools that help manage diverse sets of procedures, data sets, and intermediate data products can enable analysts to work and collaborate more effectively.

3. CHALLENGES IN THE ANALYSIS PROCESS

We identified five high-level tasks that repeatedly occurred in respondent's analysis efforts:

3.1 DISCOVERY

Throughout their work, analysts acquired data necessary to complete their tasks with large organization, finding ad understanding relevant data was often a significant bottleneck.

3.1.1 Where is my data?

For 17 analysts, finding relevant data distributed across multiple databases, database tables and/or files was very time consuming. Organizations often lacked sufficient documentation or search capabilities to enable efficient identification of desired data. Instead, analysts relied on their colleagues: they often asked database administrators or others for help. One analyst described the problem: It is really hard to know where the data is. We have all the data, but there is no huge schema where we can say this data is here and this variable is there. It may be written but the wiki is very stale: pointers don't point to the right place and it changes really fast. The best thing you can learn working here is who to ask, because in their head a lot of people know a lot of stuff. It's more like folklore. Knowledge is transmitted as you join. Some organizations also restricted access to data, requiring an appropriate administrator to grant privileges. In some cases, the administrator who set up the data may have already left the company.

3.1.2 Field Definitions

The difficulties in discovering data were compounded by the difficulty of interpreting certain fields in a database. In at least 16 instances, analysts described situations in which fields were coded and required look ups against external tables. Foreign key definitions help identify the appropriate table to perform lookups, but these definitions were often missing in relational databases and non-existent in other types of data stores. Even without coding, missing units or metadata created ambiguity. For instance, one analyst noted that many date-time fields were stored without time zone information. The analysts had torn construct time zones from corresponding geographic information. In 8 reported cases, schema drift lead to redundant columns. One company we interviewed had a database table containing four columns containing job titles for its users. These columns evolved over time, were often conflicting and there was no documentation describing which column was up-to-date or appropriate to use.

FOR EXAMPLE

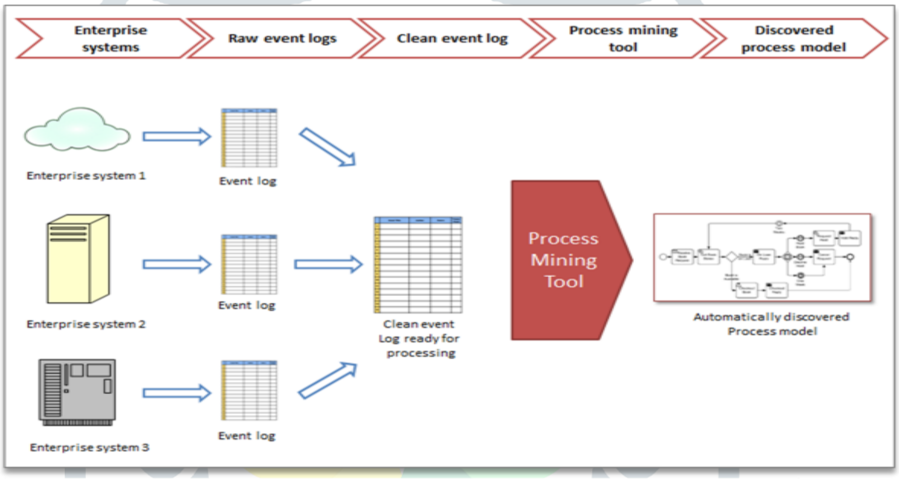


Figure 1: Process discovery from events logs

Figure 1: Focuses on process discovery. The process flows generated by process-mining tools are more reliable than those created by interviewing process stakeholders. The event logs reveal the real process, rather than how it's perceived it actors. Exception scenario us (most likely neglected in interviews) can be discovered, and overall process performance is highlighted. This process-discovery scenario provides data for future process analysis and improvement, and a basis for further analysis, such as process compliance, or a predication of process paths based on historical data.

3.2 WRANGLING

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw"" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. A data wrangler is a person who performs these transformation operations. This may include further munging, data visualization, data aggregation, training a statistical model as well as many other potential uses. Data munging as a process typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data using algorithms (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use. The data transformations are typically applied to distinct entities (e.g. fields, rows, columns, data values etc.) within a data set, and could include such actions as extractions, parsing, joining, standardizing, augmenting, cleansing ,consolidating and filtering to create desired wrangling outputs that can be leveraged downstream The recipients could be individuals, such as data architects or data scientists who will investigate the data further, business users who will consume the data directly in reports, or systems that will further process the data and write it into targets such as data warehouses, data lakes or downstream applications. Depending on the amount and format of the incoming data, data wrangling has traditionally been performed manually (e.g. via spreadsheets such as Excel) or via hand-written scripts in languages such as Python or SQL.R, a language often used in data mining and statistical data analysis, is now also often used for data wrangling. Other terms for these processes have included data franchising, data preparation and data munging.

3.2.1Data Integration

Data integration involves combining data residing in different sources and providing users with a unified view of them. This process becomes significant in a variety of situations, which include both commercial (such as when two similar companies need to merge their databases) and scientific (combining research results from different bioinformatics repositories, for example) domains. The data warehouse approach is less feasible for datasets that are frequently updated, requiring the extract, transform, load (ETL) process to be continuously re-executed for synchronization.

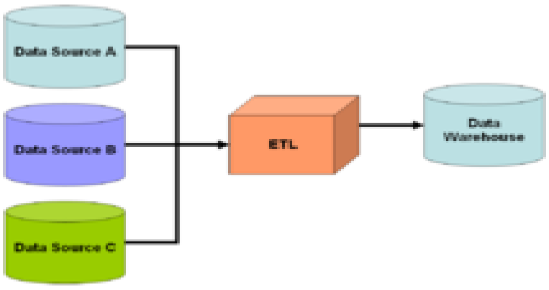


Figure 1: Simple schematic for a data warehouse. The Extract, transform, load (ETL) process extracts information from the source databases, transforms it and then loads it into the data warehouse.

Data integration favored loosening the coupling between data[1] and providing a unified query-interface to access real time data over a mediated schema (see figure 2), which allows information to be retrieved directly from original databases. This is consistent with the SOA approach popular in that era. This approach relies on mappings between the mediated schema and the schema of original sources, and transforms a query into specialized queries to match the schema of the original databases. Such mappings can be specified in 2 ways : as a mapping from entities in the mediated schema to entities in the original sources (the "Global As View" (GAV) approach), or as a mapping from entities in the original sources to the mediated schema (the "Local As View" (LAV) approach).

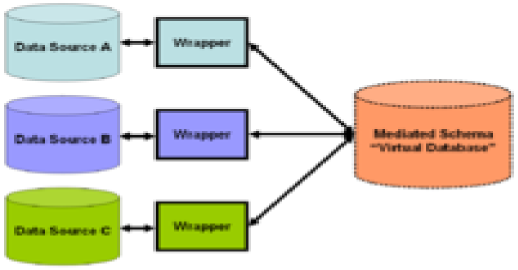


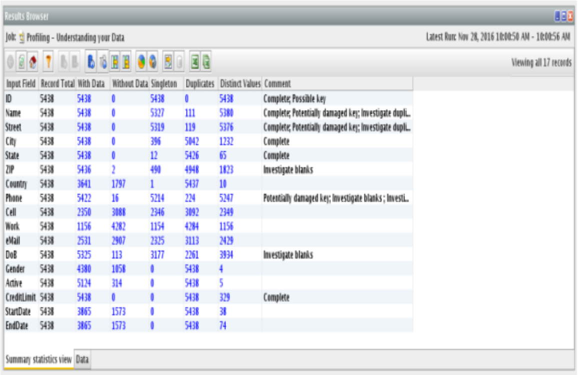
Figure 2: Simple schematic for a data-integration solution. A system designer constructs a mediated schema against which users can run queries. The virtual database interfaces with the source databases via wrapper code if required.

3.3 PROFILING

The first step in a data quality project is profiling the data. Data profiling is the analysis of data to clarify structure, content, relationships and derivation rules. Profiling helps not only to understand anomalies and to assess data quality, but also to discover, register, and asses enterprise metadata, and so the purpose of data profiling is both to validate metadata when it is available and discover metadata when it is not. EDQ comes with several built-in transformations to profile our data quickly to start discovering patterns and quality issues. Some of the most important transformations for profiling are:

3.3.1Quick stats profiler

Analyses high level completeness, duplication, and value frequency across many attributes, highlights possible issues .For example, we could use this transformation to identify duplicate values in columns that should contain unique values, or columns with more unique values than expected.



Results Browser

Job: Profiling - Understanding your Data

Latest Run: Nov 28, 2016 10:00:50 AM - 10:00:56 AM

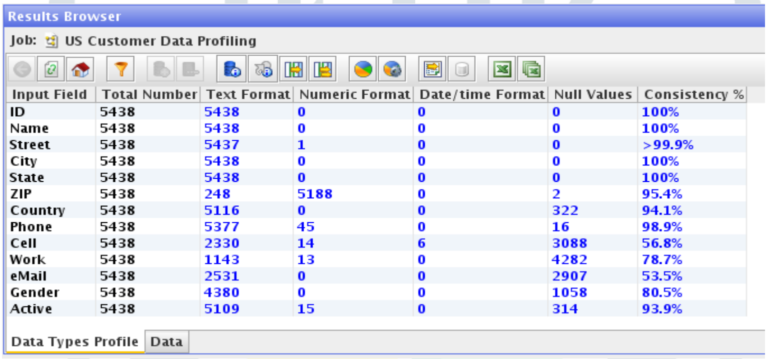
Viewing all 17 records

Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values	Comment
ID	5438	5438	0	5438	0	5438	Complete; Possible key
Name	5438	5438	0	5327	111	5380	Complete; Potentially damaged key; Investigate dupli...
Street	5438	5438	0	5319	119	5376	Complete; Potentially damaged key; Investigate dupli...
City	5438	5438	0	396	5042	1232	Complete
State	5438	5438	0	12	5426	65	Complete
ZIP	5438	5436	2	490	4948	1823	Investigate blanks
Country	5438	3641	1797	1	5437	10	
Phone	5438	5422	16	5214	224	5247	Potentially damaged key; Investigate blanks ; Investi...
Cell	5438	2350	3088	2346	3092	2349	
Work	5438	1156	4282	1154	4284	1156	
eMail	5438	2531	2907	2325	3113	2429	
DoB	5438	5325	113	3177	2261	3934	Investigate blanks
Gender	5438	4380	1058	0	5438	4	
Active	5438	5124	314	0	5438	5	
CreditLimit	5438	5438	0	0	5438	329	Complete
StartDate	5438	3865	1573	0	5438	38	
EndDate	5438	3865	1573	0	5438	74	

Figure 1: Example of the results of the Quick stats profiler in EDQ. We can start discovering issues in the data using this transformation, for example, fields with null values, or with more unique values than expected.

3.3.2 Data types profiler

Analysts attribute values for their data types, and assess data type consistency. In the example below, we can see that some columns have inconsistent data types, and we can clearly identify wrong values (the Street column has 1 numeric value and the rest text, the Cell column has 6 Date/time values, the Active columns have 15 numeric values, etc.).



Results Browser

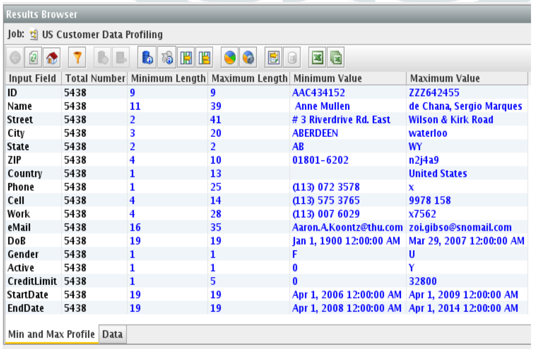
Job: US Customer Data Profiling

Input Field	Total Number	Text Format	Numeric Format	Date/time Format	Null Values	Consistency %
ID	5438	5438	0	0	0	100%
Name	5438	5438	0	0	0	100%
Street	5438	5437	1	0	0	>99.9%
City	5438	5438	0	0	0	100%
State	5438	5438	0	0	0	100%
ZIP	5438	248	5188	0	2	95.4%
Country	5438	5116	0	0	322	94.1%
Phone	5438	5377	45	0	16	98.9%
Cell	5438	2330	14	6	3088	56.8%
Work	5438	1143	13	0	4282	78.7%
eMail	5438	2531	0	0	2907	53.5%
Gender	5438	4380	0	0	1058	80.5%
Active	5438	5109	15	0	314	93.9%

Figure 2: Example of the results of the Data type's profiler in EDQ. This transformation is very useful to find data type consistency issues.

3.3.3Max/Minprofiler

Finds minimum and maximum values longest, shortest, lowest, and highest.



Results Browser

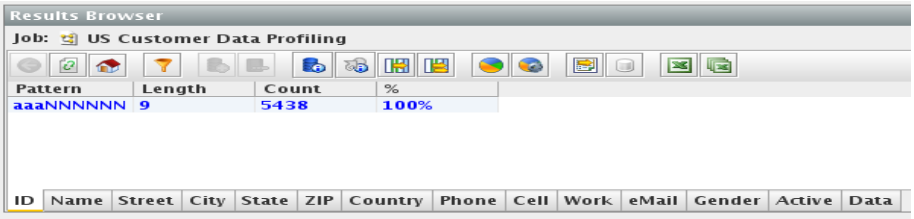
Job: US Customer Data Profiling

Input Field	Total Number	Minimum Length	Maximum Length	Minimum Value	Maximum Value
ID	5438	9	9	AAC434152	ZZZ642455
Name	5438	11	39	Anne Mullen	de Chana, Sergio Marques
Street	5438	2	41	# 3 Riverdrive Rd. East	Wilson & Kirk Road
City	5438	3	20	ABERDEEN	waterloo
State	5438	2	2	AB	WY
ZIP	5438	4	10	01801-6202	n2j4a9
Country	5438	1	13		United States
Phone	5438	1	25	(113) 072 3578	x
Cell	5438	4	14	(113) 575 3765	9978 158
Work	5438	4	28	(113) 007 6029	x7562
eMail	5438	16	35	Aaron.A.Koontz@thu.com	zoi.gibso@snomail.com
DoB	5438	19	19	Jan 1, 1900 12:00:00 AM	Mar 29, 2007 12:00:00 AM
Gender	5438	1	1	F	U
Active	5438	1	1	0	Y
CreditLimit	5438	1	5	0	32800
StartDate	5438	19	19	Apr 1, 2006 12:00:00 AM	Apr 1, 2009 12:00:00 AM
EndDate	5438	19	19	Apr 1, 2008 12:00:00 AM	Apr 1, 2014 12:00:00 AM

Figure 3: example other results of the max/min profile in EDQ. One of the basic profiles that we can do on our data is an analysis of the maximum and minimum values of each field to find wrong values.

3.3.4 Frequency profiler

Analyses value frequency across many attributes. In the image below, we can see how this transformation can be used to rapidly identify consistency problems in the values of the data.

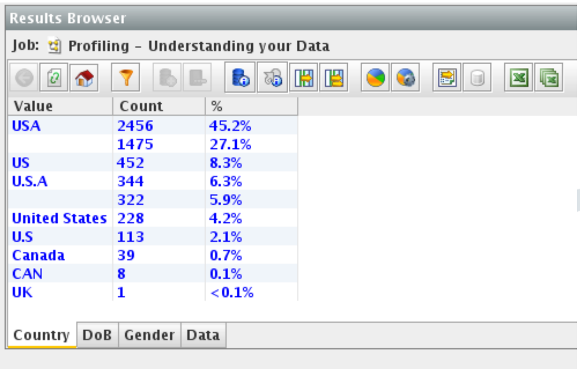


Pattern	Length	Count	%
aaaNNNNNN	9	5438	100%

Figure 4: Example of the results of the Frequency profiler in EDQ. We can already see data quality.

3.3.5 Patterns profiler

Analyses character patterns and pattern frequency across many attributes.

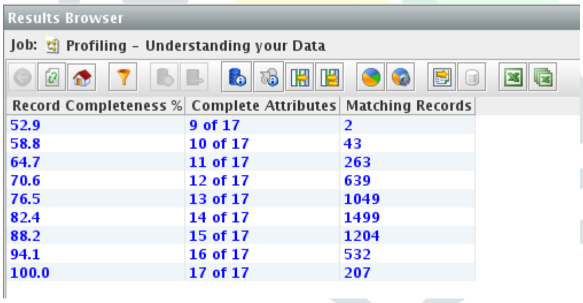


Value	Count	%
USA	2456	45.2%
	1475	27.1%
US	452	8.3%
U.S.A	344	6.3%
	322	5.9%
United States	228	4.2%
U.S	113	2.1%
Canada	39	0.7%
CAN	8	0.1%
UK	1	<0.1%

Figure 5: Example of the results of the Patterns profiler in EDQ, very useful to find data quality issues in fields such as zip codes, telephone numbers, or email addresses.

3.3.6Record completeness profiler

Analyses records for their completeness across many attributes.



Record Completeness %	Complete Attributes	Matching Records
52.9	9 of 17	2
58.8	10 of 17	43
64.7	11 of 17	263
70.6	12 of 17	639
76.5	13 of 17	1049
82.4	14 of 17	1499
88.2	15 of 17	1204
94.1	16 of 17	532
100.0	17 of 17	207

Figure 6: Example of the results of the Record completeness profiler in EDQ.

Other profiling transformations include, but are not limited to, number profiler, character profiler, date profiler, and RegEx patterns profiler.

3.4 Modeling

A data model is a description of how data should be used to meet the requirements given by the end user. Data modeling helps to understand the information requirements. Data modeling differs according to the type of the business, because the business processes or each sector is different, and it needs to be identified in the modeling stage. There are three main designs for the data model, namely conceptual design, logical design and the physical design.

There are four types of data models identified

Conceptual Data Models
Highest-level relationships between different entities.Conceptual data modeling is the most crucial stage in the database design process.

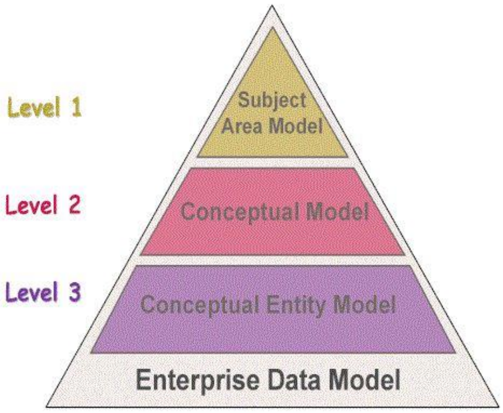


Figure 1: Conceptual Data Modeling Process

Enterprise Data Models
Addresses unique requirements of a specific business. However this is similar to conceptual data modeling. It is independent of “how” the data is physically sourced, stored, processed or accessed. The model unites, formalizes and represents the things important to an organization, as well as the rules governing them.

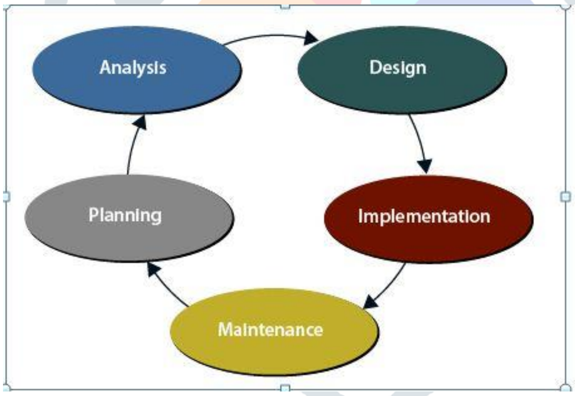
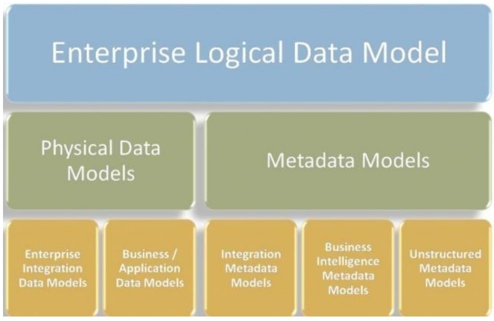


Figure 2: Enterprise Data Modeling Structure

Logical Data Modeling

Illustrates the specific entities, attributes, and relationships involved in a business function. This serves as the basis for the creation of the physical data model. Actual implementation of the conceptual model is called a logical data model. To implement one conceptual data model may require multiple logical data models. Data modeling defines the relationships between data elements and structures.


Enterprise Logical Data Model
Physical Data Models
Metadata Models
Enterprise Integration Data Models
Business / Application Data Models
Integration Metadata Models
Business Intelligence Metadata Models
Unstructured Metadata Models

Physical Data Modeling

Physical data model represents how the model will be built in the database. A physical database model shows all table structures, including column name, column data type, column constraints, primary key, foreign key, and relationships between tables.

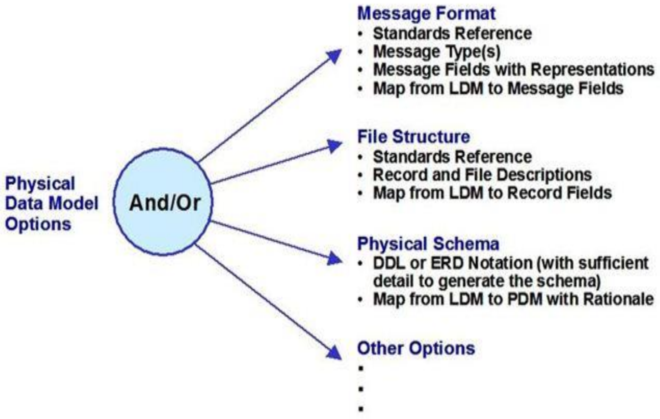

Physical Data Model Options
And/Or
Message Format
• Standards Reference
• Message Type(s)
• Message Fields with Representations
• Map from LDM to Message Fields
File Structure
• Standards Reference
• Record and File Descriptions
• Map from LDM to Record Fields
Physical Schema
• DDL or ERD Notation (with sufficient detail to generate the schema)
• Map from LDM to PDM with Rationale
Other Options
•
•
•

Figure 4: Physical Data Model

3.5 REPORTING

Analysts typically reported insights gained from modeling to other analysts or business units. The two most-cited challenges were communicating assumptions and building interactive reports.

3.5.1 Communicating Assumptions

One complaint about distributing and consuming reports (made by 17 analysts) is poor documentation of assumptions made during analysis. Analysts typically performed a sequence of operations that can affect the interpretation of results, such as correcting outliers, imputing missing data or aggregating data. These operations are often context specific, with no standards for each analysis. In other cases, analysts imposed their own definitions on underspecified concepts. One medical analyst analyzed patient episodes that correspond to all visits to a hospital to treat a given symptom. However, the database did not contain an episode identifier associated with each patient visit. The analysts had to use heuristics, such as the duration between visits, to group hospital visits into episodes. This heuristic was imprecise, as hospitals may treat a patient concurrently for two different symptoms or for the same symptom after a long period of time. Analysts often lost track of all the operations they performed and their rationale for performing them. Even when assumptions were tracked, they were often treated as foot notes instead of first-class results. One analyst cited that his boss often looked at summary charts without reading the fine print. For instance, an average calculated from three data points would be marked with an asterisk that was then regularly overlooked.

3.5.2 Static Reports

A number of analysts (17/35) also complained that reports were too inflexible and did not allow interactive verification or sensitivity analysis. Often reporting and charting tools were used directly on the output data and contained no knowledge of how the original input data was filtered, transformed or modeled. Much of this work was done before output data was loaded into the tool. Because reporting tools have no access to data provenance, it was often impossible to modify parameters or assumptions to see how the conclusions would change. Viewers can not verify questions such as “how might user acquisition rates change if more money was spent on marketing?”

4. CONCLUSION

This paper presented by the five high level challenges faced by data analysis .The challenges process of intelligence and enterprise, these analysts often work on different types of data to the implement of the task in enterprise data analysis and visual analytic tools .To improve the quality of data analysis take place in enterprise. These challenges are typically more acute within large organization with a diverse and distributed set of data source.

REFERENCES

[1] SeanKandel, Andreas Paepcke, Joseph M.Hellerstein, and Jeffrey Heer. Enterprise data analysis and visualization: An Interview study. Manuscript received 31 March 2012; accepted| August 2012; posted online 4 Oct 2012; mailed on 5 Oct 2012.

[2] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In Proc. IEEE Information Visualization (InfoVis), pages 111–117, 2005.

[3] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo. VisTrails: visualization meets data management. In Proc. ACM SIGMOD, pages 745–747, 2006.

[4] G. Chin, O. A. Kuchar, and K. E.Wolf.Exploring the analytical processes of intelligence analysts. In Proc. ACM Human Factors in ComputingSystems (CHI), pages 11–20, 2009.

[5] P. Isenberg, D. Fisher, M. Morris, K. Inkpen, and M. Czerwinski. Anexploratory study of co-located collaborative visual analytics around atabletop display. In Proc. IEEE Visual Analytics Science and Technology(VAST), pages 179–186, 2010.

[6] Y. Kang and J. Stasko.Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In Proc. IEEE Visual Analytics Science and Technology (VAST), pages 21–30, 2011.

[7] X. Jiang and J. Zhang.text visualization method for cross-domainresearch topic mining. Journal of Visualization, pp. 1–16, 2016.doi: 10.1007/s12650-015-0323-9.

JETIR