International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 5

System Model for Processing on Multi-Format of Dataset

Archana H M[1], Tejaswini Busnur[3], Dr. Poornima B[2]
[1]Student, Department of Computer Science and Engineering
[2]Head of Department of Information Science and Engineering
[1,2]Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India
[3]Associate Consultant, Capgemini Pvt Ltd, India

ABSTRACT

The problem associated with Big Data is having following feature are called 3V features: volume: large amount of data, velocity: data processing rate and variety: collection of structured data, semi-structured data, and unstructured data, the three V's of data that has arrived in unprecedented ways. In the Present years there are many sources of data form, where we obtain variety of data of same domain for processing, when that data become huge to handle we require efficient system to handle that data and to process that data for query prediction or result prediction. The 3V highlights represent a stupendous test to conventional information processing systems since these frameworks either can't scale to the tremendous data volume in a survey way or neglect to deal with information with assortment of types. This undertaking presents another framework called system model for big data processing on multi-variety of dataset to handle the Big Data's information assorted variety of challenges. The significant commitment of this work is an engineering plan that empowers clients to process multi-structured datasets in a solitary framework.

I. INTRODUCTION

Machine Learning is researching concept which is growing in present market. Machine learning plays vital role in all data analyzing paradigm. Among all data analyzation such as predicting present trend in twitter and face book, predicting diseases by analyzing patients data's etc.., all these concept relays on machine learning.

The problem associated with Big Data is having following feature of so called 3V features: volume: large amount of data, velocity: data processing rate and variety: collection of structured data, semi-structured data, and unstructured data, the three V's of data that has arrived in unprecedented ways.

Present years there are many sources of data from we get many data of same domain for processing when that data become more to handle we required efficient system to handle that data and to process that data for query prediction or result prediction.

II. RELATED WORK

Large amount of data handling frameworks can be categorized into the following: a) Parallel Databases, c) Map Reduce based frameworks, 3)DAG based information preparing frameworks, d) Actor-like frameworks and e)hybrid frameworks. An extensive study could be discovered in, and another benchmark called Big Bench, was additionally as of late proposed to assess and look at the execution of various huge data processing frameworks.

Parallel databases are chiefly intended for handling structured data sets where every datum (called a record) entirely shapes a relational schema. These frameworks utilize information parceling and partitioned execution systems for elite question handling.

Late parallel database frameworks additionally utilize the column oriented processing system to try and enhance the execution of analytical workloads, for example, OLAP questions. Parallel data bases are used to process huge amount of data from peta-byte to tera-byte of dataset but cost of equipmenting and programming is very high. The main disadvantage of parallel databases is that those frameworks can't able to process unstructured information. Because of some of issues related to parallel data base processing, by survey Hadoop can efficiently process unstructured information.

Our system model for processing big data on multi-variety of dataset, then again, has been outlined and assembled from scratch to give the adaptability, productivity and adaptability discovered in both stages.

Dean and Ghemawat proposes Map Reduce concept. The first Map Reduce system was created to build some index for extensive web corpus. Be that as it may, the capacity of utilizing Map Reduce as a general information investigation device for preparing both organized data and unstructured information was immediately perceived. Map Reduce gains notoriety because of its straightforwardness and adaptability. Even though the programming model is moderately basic (just consists of two capacities), clients, in any case, can indicate any sorts of computations in the guide() and diminish() usage. Map Reduce is likewise amazingly adaptable and strong to process slave workers. The main drawback of Map Reduce is to processing structured (social) information and chart information. Number of experimental work has been proposed to improve the execution of Map Reduce based on handling relational information. The main goal of present work is to focus on processing structured dataset by using Map Reduce programming model.

System model for processing big data on multi-variety of dataset proposed to use different information models to process different data and utilize a typical simultaneous programming model for parallel information processing. Dryad processing model is one of the research areas at Microsoft Corporation. Proposed work is not same as that of Dryad model, in present work mainly focuses on simultaneous programming model.

The concept of actor like programming model is developed to simplify the process of concurrent programming. Proposed work is mainly focuses on simultaneous programming model is inspired by the actor display. Be that as it may, unique in relation to Storm and S4, our proposed system is intended for cluster information preparing. In any case, occupations of Storm and S4 may never end. Start is likewise an information investigative framework in light of Actor programming model. The framework presents a deliberation called resilient disseminated dataset (RDD) for adaptation to non-critical failure. In Spark, the input and yield of every administrator must be a RDD. RDD is further implemented as an in-memory information structure for better recovering execution.

Our approach is unique in relation to Spark in that unit is independent of the underline stockpiling framework. Clients can perform in memory information systematic assignments by stacking information into certain in memory storage frameworks. Besides, not quite the same as RDD, epic utilizes a mixture plan to accomplish adaptation to internal failure.

III. OVERVIEW OF SYSTEM

A. MODULE DESCRIPTION

Structured Data Module: Current information distribution center contains organized information and just organized information. It means a data which represent particular entity data such as empname, empno, salary yet we realize that the empno esteem runs with a particular individual, thus organized. Structured module procedures, examinations and gives about on the organized information. In this venture, Transaction dataset of E-business site is taken and we are making a recommender framework.

Semi-Structured Data Module: Here data is going to represent the particular information but not in proper format, where data content different attribute information, but not in organized and not as un-organized, it contain some special element to separate data attributes. This module processes, analyses and gives results on the semi-structured data. In this project, details of different abalone are taken by system and it predicts the age of all the abalone.

Unstructured Data Module: In unstructured data format their will all the data regarding the element but not been particularly organized is viewed as unstructured. The rundown of really unstructured information incorporates free content, for example, records created in organization, pictures and recordings, sound documents, and a few kinds of internet based life. On the off chance that the protest be put away conveys no labels (metadata about the information) and has no settled construction, steady association it is unstructured. Nonetheless, in an indistinguishable classification from unstructured information there are numerous kinds of information that do have at any rate some association. This module procedures, investigations and gives result on unstructured information. In this task, patient's restorative data is taken and then system predicts he/she has coronary illness or not.

B. SYSTEM ARCHITECTURE

Figure 1 shows the architectural design of the system which presents a framework called system model for processing big data on multi-variety of data set to handle the Big Data's information assorted variety challenge. The significant commitment of this work is an engineering plan that empowers clients to process multi-structured datasets in a solitary framework. We found that albeit distinctive frameworks intended for different types of information, they all offer the same system model and break down entire calculation into free calculations for parallelization. This new model going to process all type of formatted and unformatted data which is arrived in real world.

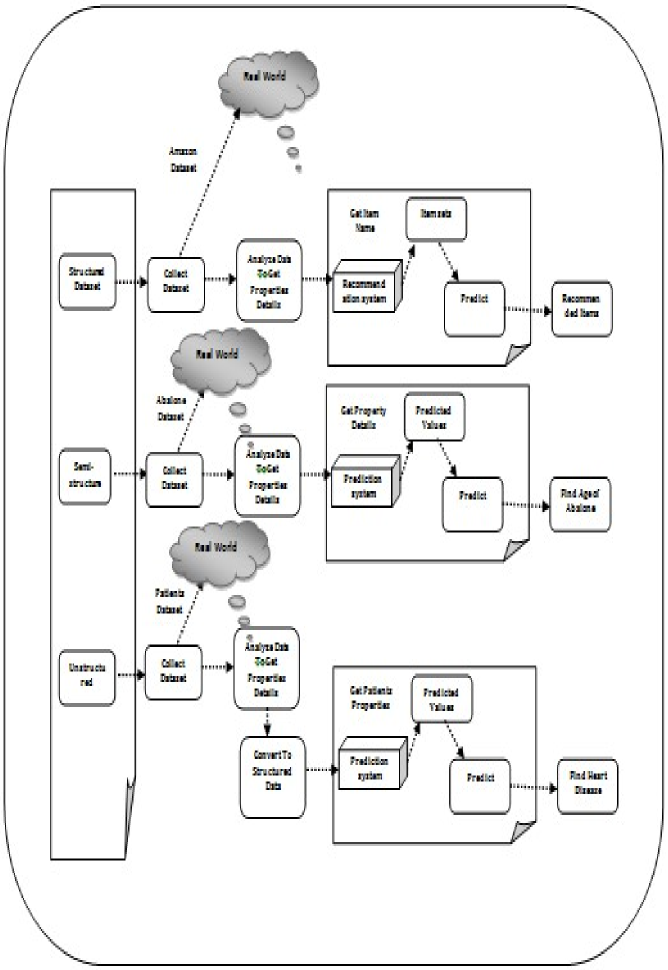


Figure 1: Architectural design of system

Advantages: To handle the information assortment challenge, the cutting edge approach supports a crossover design.

This approach utilizes a half breed framework to handle multi-organized datasets (i.e., assortment of information writes: organized information, content, chart, unstructured information).

The multi-organized dataset is put away in an assortment of frameworks in light of writes (e.g., organized datasets are put away in a data base; unstructured datasets are put away in Hadoop). Finally, the yield of those sub-employments will be Stacked into a solitary framework (Hadoop or data base) with appropriate information development to create the last out comes.

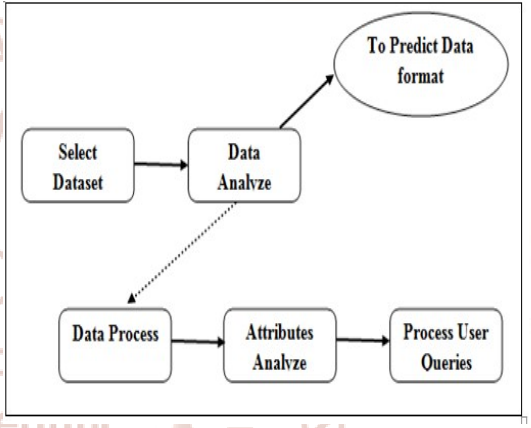


Figure 2: Steps in Processing Multi-Format of Dataset

Methodology: Figure 2 shows steps in processing multi-format of dataset, the steps are as follows:

Step1: Select Dataset: First select dataset which is stored in media for processing.

Step2: Data Analyze: After selecting dataset, first Identify the format of Data set whether it is a Structured, Semi-structured or Un-structured.

Step3: Data Process: After identification of dataset format then start finding attribute and their properties according to application to analyze for prediction.

Step4: Attribute Analyze: After data process we obtain the information regarding attributes and their relationship and then find out the important attribute element in the given dataset.

Step5: Process the user queries: Finally it is important to process user queries on these analyzed dataset to find result.

IV. IMPLEMENTATION

Implementation of any software has some constraints regarding cost, effort to learn and use. The application needs minimal key stroke and more use of mouse and application is secured, robust, and easily customizable and personalize.

Structured Module: Get dataset file and analyze. Select the dataset from database and analyze the dataset and store the analyzed dataset in database and then find the average and give rating to unrated items, finally get recommended item: Select the item and view the review ratings and apply for recommendation and get the best recommended item.

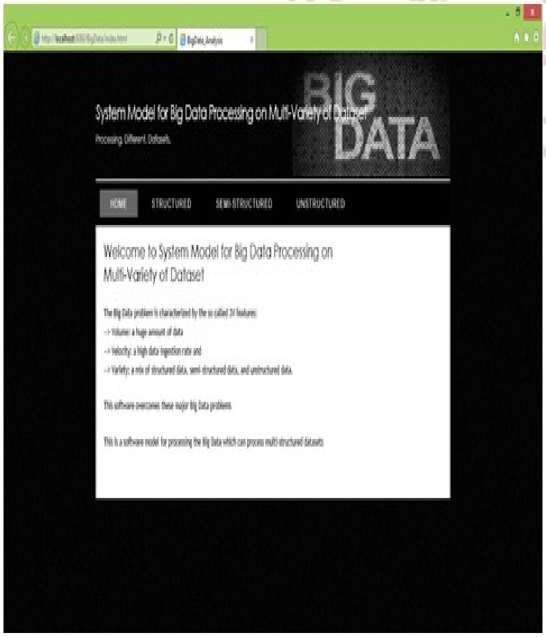
Semi-Structured Module: Get dataset file and analyze. Select the dataset from database and analyze the dataset and store the analyzed dataset in database, finally predict the age of abalone by computing attribute information.

Unstructured Module: Get dataset file and analyze. Select the dataset from database and analyze the dataset and store the Analyzed dataset in database, finally predict the heart disease by computing patient's attribute information.

V. RESULTS

Finally we designed and developed a system model for processing multi-format of dataset.

Figure 3 shows home page of system model for processing big data on multi-variety of dataset, this system model includes three sub-modules they are as follows structured, semi-structured and unstructured.



System Model for Big Data Processing on Multi-Variety of Dataset

Processing. Different. Datasets.

Welcome to System Model for Big Data Processing on Multi-Variety of Dataset

The Big Data problem is characterized by the so called 3V features:

--> Volume: a huge amount of data

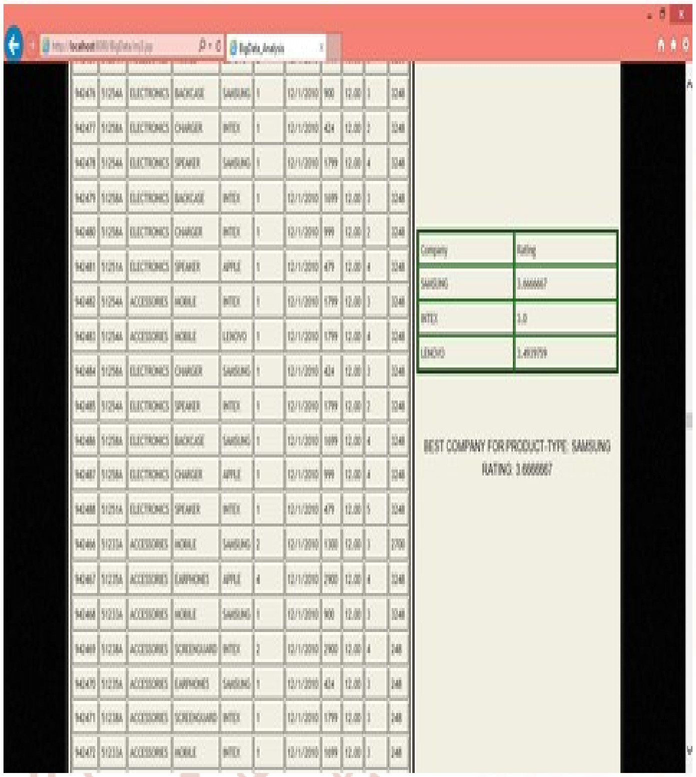
--> Velocity: a high data ingestion rate and

--> Variety: a mix of structured data, semi-structured data, and unstructured data.

This software overcomes these major Big Data problems

This is a software model for processing the Big Data which can process multi-structured datasets

Figure 3: System Home Page



942476	51254A	ELECTRONICS	BACKCASE	SAMSUNG	1	12/1/2010	900	12.00	3	3248
942477	51258A	ELECTRONICS	CHARGER	INTEX	1	12/1/2010	424	12.00	2	3248
942478	51254A	ELECTRONICS	SPEAKER	SAMSUNG	1	12/1/2010	1799	12.00	4	3248
942479	51258A	ELECTRONICS	BACKCASE	INTEX	1	12/1/2010	1699	12.00	3	3248
942480	51258A	ELECTRONICS	CHARGER	INTEX	1	12/1/2010	999	12.00	2	3248
942481	51251A	ELECTRONICS	SPEAKER	APPLE	1	12/1/2010	479	12.00	4	3248
942482	51254A	ACCESSORIES	MOBILE	INTEX	1	12/1/2010	1799	12.00	3	3248
942483	51254A	ACCESSORIES	MOBILE	LENOVO	1	12/1/2010	1799	12.00	4	3248
942484	51258A	ELECTRONICS	CHARGER	SAMSUNG	1	12/1/2010	424	12.00	3	3248
942485	51254A	ELECTRONICS	SPEAKER	INTEX	1	12/1/2010	1799	12.00	2	3248
942486	51258A	ELECTRONICS	BACKCASE	SAMSUNG	1	12/1/2010	1699	12.00	4	3248
942487	51258A	ELECTRONICS	CHARGER	APPLE	1	12/1/2010	999	12.00	4	3248
942488	51251A	ELECTRONICS	SPEAKER	INTEX	1	12/1/2010	479	12.00	5	3248
942466	51233A	ACCESSORIES	MOBILE	SAMSUNG	2	12/1/2010	1300	12.00	3	2700
942467	51235A	ACCESSORIES	EARPHONES	APPLE	4	12/1/2010	2900	12.00	4	3248
942468	51233A	ACCESSORIES	MOBILE	SAMSUNG	1	12/1/2010	900	12.00	3	3248
942469	51238A	ACCESSORIES	SCREENGUARD	INTEX	2	12/1/2010	2900	12.00	4	248
942470	51235A	ACCESSORIES	EARPHONES	SAMSUNG	1	12/1/2010	424	12.00	3	248
942471	51238A	ACCESSORIES	SCREENGUARD	INTEX	1	12/1/2010	1799	12.00	3	248
942472	51233A	ACCESSORIES	MOBILE	INTEX	1	12/1/2010	1699	12.00	3	248

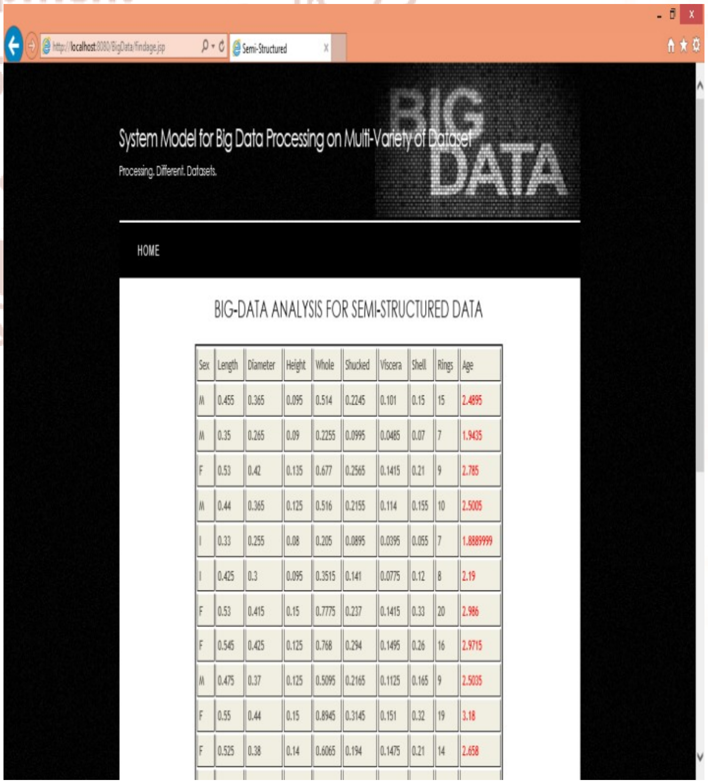
Company	Rating
SAMSUNG	3.666667
INTEX	3.0
LENOVO	3.493759

BEST COMPANY FOR PRODUCT-TYPE: SAMSUNG
RATING: 3.666667

Figure 4: Result Analysis in Structured Dataset

Figure 4 shows result analysis-3, displays company name to purchase product with best rating.

Figure 5 shows result analysis in Semi-Structured dataset, provide query result after analyzing suitable attribute information.



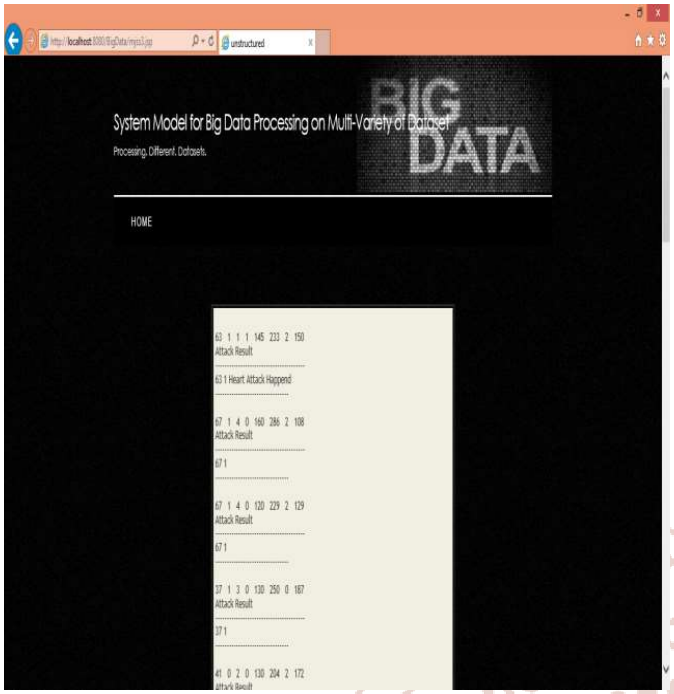
System Model for Big Data Processing on Multi-Variety of Dataset

Processing. Different. Datasets.

BIG-DATA ANALYSIS FOR SEMI-STRUCTURED DATA

Sex	Length	Diameter	Height	Whole	Shucked	Viscera	Shell	Rings	Age
M	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	2.4895
M	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	1.9435
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	2.785
M	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	2.5005
I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1.8889999
I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	2.19
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	2.986
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	2.9715
M	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	2.5035
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	3.18
F	0.525	0.38	0.14	0.6065	0.194	0.1475	0.21	14	2.658

Figure 5: Result Analysis in Semi-Structured dataset



System Model for Big Data Processing on Multi-Variety of Dataset

Processing. Different. Datasets.

```
63 1 1 1 145 233 2 150
Attack Result
-----------------------
63 1 Heart Attack Happend
-----------------------
67 1 4 0 160 286 2 108
Attack Result
-----------------------
67 1
-----------------------
67 1 4 0 120 229 2 129
Attack Result
-----------------------
67 1
-----------------------
37 1 3 0 130 250 0 187
Attack Result
-----------------------
37 1
-----------------------
41 0 2 0 130 204 2 172
Attack Result
```

Figure 6: Result Analysis in Unstructured Dataset

Figure 6 shows result analysis in unstructured dataset, provide final query result after complete processing of unstructured dataset.

VI. CONCLUSION

As we said the features of Big Data's , data volume, velocity and variety-the three V's-of data that has arrived in unprecedented ways. Present years there are many sources of data from we get many data of same domain for processing when that data become more to handle we required efficient system to handle that data and to process that data for query prediction or result prediction. The 3V highlights represent a stupendous test to conventional information processing systems since these frameworks either can't scale to the tremendous data volume in a survey way or neglect to deal with information with assortment of types. This undertaking presents another framework called system model for big data processing on multi-variety of dataset to handle the Big Data's information assorted variety challenge. The significant commitment of this work is an engineering plan that empowers clients to process multi-structured datasets in a solitary framework.

REFERENCES

1. The hadoop offical website. http://hadoop.apache.org/.
2. The storm project offical website. http://storm-project.net/.
3. D. Jiang, B. C. Ooi, L. Shi, and S. Wu. The performance of Map Reduce: an in-depth study. *PVLDB*, 3(1-2), Sept. 2010.
4. A. Abouzeid, K. Bajda-Pawlikowski, D. Abadi, A. Silberschatz, and A. Rasin. Hadoop D B: an architecturalhybrid of Map Reduce and dbms technologies for analytical workloads. *PVLDB*, 2(1), Aug. 2009.
5. S. Fushimi, M. Kitsuregawa, and H. Tanaka. An overview of the system software of a parallel relational database machine grace. In *VLDB*, 1986.
6. J. Dean and S. Ghemawat. Map Reduce: simplified data processing on large clusters. *Commun. ACM*, 51(1), Jan. 2008.
7. M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: distributed data-parallel programs from sequential building blocks. *SIGOPS Oper. Syst. Rev.*, 41(3), Mar. 2007.
8. D. J. DeWitt, A. Halverson, R. Nehme, S. Shankar, J. Aguilar-Saborit, A. Avanes, M. Flasza, and J. Gramling. Split query processing in poly base. In *SIGMOD*, 2013.
9. M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoic a. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *NSDI'12*, 2012.
10. Y. Bu, B. Howe, M. Balazinska, and M. D. Ernst. HaLoop: efficient iterative data processing on large clusters. *VLDB*, 3(1-2), Sept. 2010.
11. S. Fushimi, M. Kitsuregawa, and H. Tanaka. An overview of the system software of a parallel relational database machine grace. In VLDB, 1986.
12. A. Ghazal, T. Rabl, M. Hu, F. Raab, M. Poess, A. Crolotte, and H.-A. Jacobsen. Big Bench: Towards an industry standard benchmark for big data analytics. In SIGMOD, 2013.
13. C. Hewitt, P. Bishop, and R. Steiger. A universal modular actor formalism for artificial intelligence. In IJCAI, 1973.
14. D. Jiang, B. C. Ooi, L. Shi, and S. Wu. The performance of Map Reduce: an in-depth study. PVLDB, 3(1-2), Sept. 2010.
15. D. Jiang, A. K. H. Tung, and G. Chen. MAP-JOIN-REDUCE: Toward scalable and efficient

data analysis on large clusters. IEEE TKDE, 23(9), Sept. 2011.

16. F. Li, B. C. Ooi, M. T. ¨Ozsu, and S. Wu. Distributed data management using Map Reduce.ACM Comput. Surv. (to appear), 46 (3), 2014.
17. G. Malewicz, M. H. Austern, A. J. C. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski. Pregel: a system for large-scale graph processing. In SIGMOD, 2010.
18. M. Stonebraker, D. J. Abadi, A. Batkin, X. Chen, M. Cherniack, M. Ferreira, E. Lau, A. Lin, S. Madden, E. O'Neil, P. O'Neil, A. Rasin, N. Tran, and S. Zdonik. C-store: a column-oriented dbms. In VLDB, 2005.
19. X. Su and G. Swart. Oracle in-database Hadoop: when Map Reduce meets RDBMS. In SIGMOD, 2012.
20. A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy. Hive: a warehousing solution over a map-reduce framework. VLDB, 2(2), Aug. 2009.
21. S. Wu, F. Li, S. Mehrotra, and B. C. Ooi. Query optimization for massively parallel data processing. In SoCC, 2011.
22. H. Yang, A. Dasdan, R. Hsiao, and D. S. Parker. Map-Reduce-Merge: simplified relational data processing on large clusters. In SIGMOD, 2007.