A Study on Big Data Techniques and Applications

K. Radha, B. Thirumala Rao
CSE, KL University, Guntur, Andhra Pradesh, India

Article Info

Article history:

Received Mar 12, 2016
Revised May 8, 2016
Accepted May 19, 2016

Keyword:

Big Data
Data-Intensive Applications
Map Reduce
Variety
Velocity
Volume

ABSTRACT

We are living in on-Demand Digital Universe with data spread by users and organizations at a very high rate. This data is categorized as Big Data because of its Variety, Velocity, Veracity and Volume. This data is again classified into unstructured, semi-structured and structured. Large datasets require special processing systems; it is a unique challenge for academicians and researchers. Map Reduce jobs use efficient data processing techniques which are applied in every phases of Map Reduce such as Mapping, Combining, Shuffling, Indexing, Grouping and Reducing. Big Data has essential characteristics as follows Variety, Volume and Velocity, Viscosity, Virality. Big Data is one of the current and future research frontiers. In many areas Big Data is changed such as public administration, scientific research, business, The Financial Services Industry, Automotive Industry, Supply Chain, Logistics, and Industrial Engineering, Retail, Entertainment, etc. Other Big Data applications are exist in atmospheric science, astronomy, medicine, biologic, biogeochemistry, genomics and interdisciplinary and complex researches. This paper is presents the Essential Characteristics of Big Data Applications and State of-the-art tools and techniques to handle data-intensive applications and also building index for web pages available online and see how Map and Reduce functions can be executed by considering input as a set of documents

Corresponding Author:

K. Radha,
CSE,
KL University,
Andhra Pradesh, India.
Email: radha.saitej@gmail.com

1. INTRODUCTION

We are living in on-Demand Digital Universe with data spread by users and organizations at a very high rate. This data is categorized as Big Data because of its Variety, Velocity, Veracity and Volume. In a heterogonous environment, this data is again divided into unstructured, semi-structured and structured. To manage Big Data, such kind of data is difficult for the present computing infrastructure. Conventional data management, analysis systems and warehousing fall short of tools to analyze this data. This data is stored in distributed file system due to its specific nature of Big Data. To store and manage Big Data Hadoop and HDFS by Apache is widely used. Analysis of Big Data is critical task as it involves large distributed file systems which would be scalable, fault tolerant and flexible. For the efficient analysis of Big Data Map Reduce is used widely. In Map Reduce for graph search clustering and classification techniques were used and also some other conventional DBMS techniques such as Joins and Indexing are used. For conventional analysis of data analyzing big data is tedious task and management tools because of its velocity, heterogeneity and volume of big data. Map Reduce overcome the problem of analyzing large distributed data sets [2], [3]. Large datasets require special processing systems; it is a unique challenge for academicians and researchers. Map Reduce jobs use efficient data processing techniques which are applied in every phases of Map Reduce such as Mapping, Combining, Shuffling, Indexing, Grouping and Reducing. Google’s technical response to the challenges of analysis and Web-scale data management was simple, by database standards, but kicked off what has become the modern “Big Data” revolution in the systems world. To handle the challenge of Web-scale storage, the Google File System was created. Google file system provides clients with the familiar operating system level byte-stream abstraction, but it does so for extremely large files whose content can span hundreds of machines in shared-nothing clusters created using inexpensive commodity hardware. Map Reduce programming framework was developed by Google to handle the challenge of processing the data in massive big files. This paradigm is described as “parallel programming for dummies” enabled Google’s developers to process the massive collections of data by writing two user-defined functions such as Map and Reduce, that the Map Reduce framework applies to the instances (map) and sorted groups of instances that share a common key (reduce) – similar to the sort of partitioned parallelism utilized in shared-nothing parallel query processing. Conventionally, Big Data is described as data is too big for existing systems to process. Big Data has essential characteristics as follows Variety, Volume and Velocity as shown in Fig.1. In a Distributed Systems world, Big Data started to become a major challenge in the late 1990’s due to the impact of world-wide web. Database technology (including parallel databases) was considered for the task, but was found to be neither well-suited nor cost-effective for those purposes. The turn of the millennium then brought further challenges as companies began to use information such as the topology of the Web and users‟ search histories in order to provide increasingly useful search results, as well as more effectively-targeted advertising to display alongside and fund those results.

The necessity to process massive quantities of data has never been greater. Not only terabyte and petabyte scale datasets rapidly becoming common place, but there is consensus that great value lies buried in them, waiting to be unlocked by the right computational tools. In the commercial world, business intelligence gathers the data from array of sources. Big Data analysis tools like Map Reduce over Hadoop, HDFS, to assist to organizations better understand their market place and customers hopefully leading to better business decisions and competitive benefits. For engineers building information processing tools and applications, large and heterogeneous datasets which are generating continuous flow of data, lead to more effective algorithms for a wide range of tasks.

2. BIG DATA CHARACTERISTICS

From Big Data [4]. There are different explanations for Big Data from 3 V’s to 4 V’s. According to Doug Laney, Volume, Velocity and Variety referred to as 3Vs [9]. According to other people special requirements, they are extended another V. The fourth V is Value, Variability [10]. Big Data is a collection of very huge data sets with diversification of types such that, it becomes tedious to process by using the state-of-the-art data processing approaches or conventional data processing platforms. In the year 2012, Gartner defined Big Data as “Big Data is High Velocity, High Volume and High variety information assets require new forms of processing to enable enhanced decision making, process optimization and insight discovery [6].

2.1. Volume

Volume is described as the relative size of the data to the processing capability. Every day we are creating 2.5 quintillion bytes of data [5]. This data is generated from everywhere such as from sensors, social media sites, digital pictures videos, purchase transaction records, etc. to overcome this volume problem requires technologies that store massive amounts of data in a scalable manner and provide distributed approaches to find that data. Apache Hadoop based solutions and massively parallel processing databases such as EMC Green plum, Calpont, EXASOL, IBM Netwzza, Teradata Kick fire.

2.2. Velocity

Velocity is described as a frequency at which the data is generated, shared and captured. The growth in sensor data from devices, and web based click stream analysis now creates requirements for greater real-time use cases. The velocity of massive data streams power the ability to parse text, identifying new patterns and detect sentiment. Key technologies that address velocity include streaming processing and complex event processing. When relational approaches no longer make sense, NoSQL databases are used. In addition to that, columnar databases, the use of in-memory data bases (IMDB), and key value stores help improve retrieval of pre-calculated data.

2.3. Variety

Spread of data types from machine to machine, social and mobile sources add new data types to conventional transactional data. Data no longer fits into neat, easy to consume structures. New types include geo-spatial, content, hardware data points, log data, machine data, mobile, physical data points, process, metrics, RFID’s search, social, web, sentiment streaming data and text. Unstructured data such as text, speech and language increasingly complicate the ability to categorize data. Some of the technologies that are dealing with unstructured data include text analytics, data mining and noisy text analytics.

3. RESEARCH METHOD

3.1. Map Reduce

1) Thus the MapReduce framework transforms a list of (key, value) pairs into a list of values.
2) These behaviors is different from the functional programming map and reduce combination, which accepts a list of arbitrary values and returns one single value that combines all the values returned by map.
3) It is necessary but not sufficient to have implementations of the map and reduce abstractions in order to implement MapReduce.
4) Distributed implementations of MapReduce require a means of connecting the processes performing the Map and Reduce phases.
5) This may be a Distributed file system.
6) Other options are possible, such as direct streaming from mappers to reducers, or for the mapping processors to serve up their results to reducers that query them.

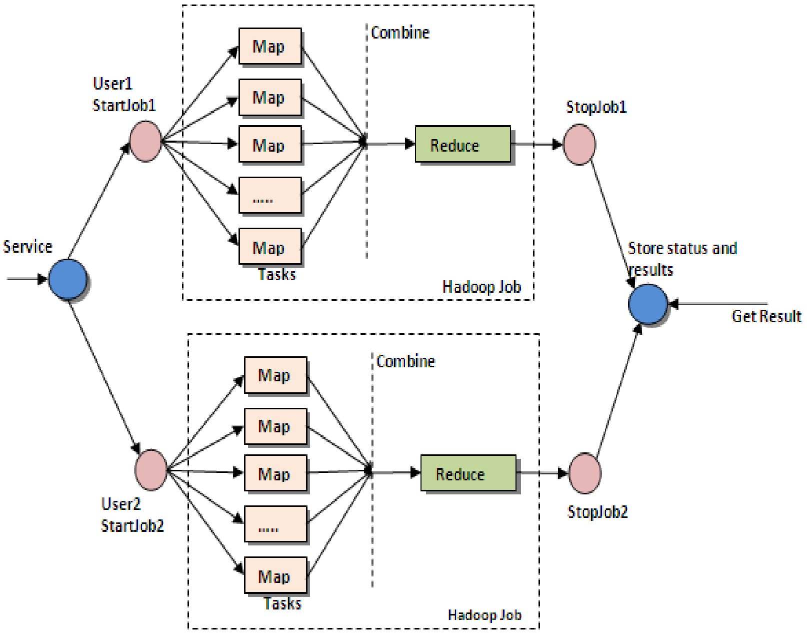

Figure 1. Work flow of Map Reduce

3.1.1. User Program

1) Typically Execution of a program begins with the user program
2) Map Reduce libraries are imported into the program and that program is splitted into the operations that are to be performed on the input dataset.
3) In a cluster every machine has a separate instance of the mapper program running on it.
4) There are master and workers. One of the copies of the program is Master and Remaining programs are assigned to work under the master called as Worker. There are M number of tasks and N Number of reduce operations to perform. The mapper picks the unused workers and assigns each of them a map task or reduce task.

3.1.2. Map Workers

1) The worker that is assigned the Map task takes the splitted input data and produces the key/value pair for every segment of input data.
2) User-defined map function is invoked by the worker node.

3) The resultant values of the Map function are buffered in the memory. Temporary data later written to the disk.
4) The physical address of these contents is passed to the Master.
5) To perform the Reduce task the master find the passes these physical memory addresses to them.

3.1.3. Reduce Workers

1) Reduce worker notified by the user's remote procedure calls to access the buffered data from the Map workers.
2) Whenever reduce worker has read all the intermediate data, it groups together all the data of the same intermediate key.
3) Various different keys map to the same task because of the parallel processing nature of the tasks. Such that the above mentioned sorting step is needed.
4) For every user every unique key and its data are passed by the reduce worker to the Reduce function.
5) Output of a Reduce task is written to an output usually to distributed file system.

3.1.4. Return to the User Program

1) After running all the Map and Reduce have been run, The Master node sends control back to the user side.
2) There are many output files available to the user as there were Reduce calls. upon completion of above mentioned set of tasks
3) These files may reinsert into another Map Reduce tasks session or they may deal as inputs for distributed processing applications.

3.2. Logical View

1) For both the Map and Reduce functions of Map Reduce, Data is assumed to be structured in (key, value) pairs.
2) Map takes one pair of data with a type in one data domain, and returns a list of pairs in a different domain: Map(k1,v1) -> list(k2,v2)
3) By applying this map function in parallel to every item in the input dataset, Parallel processing is introduced
4) It produces a list of (K2, v2) pairs for each call. After that, the Map Reduce framework collects all pairs with the same key from all lists and groups them together, thus creating one group for each one of the different generated keys.
5) This phase optimizes the input for reduce function.
6) The Reduce function is then applied in parallel to each group, which in turn produces a collection of values in the same domain: Reduce (k2, list (v2)) -> list(v3)

Algorithm
Input: Data in the form of (key, value) pairs
Output: List of data items
Algorithm:

```
1. Map data from one domain to another [Map(m1,v1) -> list(m2,v2)]
2. Optimize input for Reduce function
3. Reduce the data into more meaningful data in the same domain [Reduce(m2, list (v2)) -> list(v3)]
```

The Map and Reduce functions are necessary but not sufficient for MapReduce framework. These two functions bring the parallel processing to the algorithm as they can be executed simultaneously for each given data.

Let us take an example of building index for web pages available online and see how Map and Reduce functions can be executed. Input can be considered as a set of documents

Pseudo code for Map:

```
For each word mk in document
count(mk) = count(mk) + 1
```

Pseudo code for Reduce:

```
For each word wk over all documents
index(mk) = Sum(count(mk))
```

4. BIG DATA APPLICATIONS

Big Data is one of the current and future research boundaries. Gartner listed the “For the Next Five Years Top 10 Critical Tech Trends” [7] and “For 2013 Top 10 Strategic Technology Trends” [8]. In many areas Big Data is changed such as public administration, scientific research, business, The Financial Services Industry, Automotive Industry, Supply Chain, Logistics, and Industrial Engineering, Retail, Entertainment, etc. Other Big Data applications are exist in atmospheric science, astronomy, medicine, biologic, biogeochemistry, genomics and interdisciplinary and complex researches. Web-based applications are encounter big data such as social computing (includes online communities, reputation systems, social network analysis, prediction markets, recommender systems, Internet search indexing, Internet text and documents. There are various sensors available around us, they will generate seamless sensor data that need to be utilized for example intelligent transportation systems (ITS) [11] are based on the analysis of massive volume of complex sensor data. Data-intensive applications are large scale e-commerce [12]. This data-intensive application consists of massive number of transactions and customers. In the following subsections we will briefly introduce various applications of the Big Data problems in business, society administration and scientific research fields.

4.1. Big Data in Society Administration

Public administration has Big Data problems [14].usually population of one country is very large. In each age level require distinct public services. For instance, adults and kids require more education and elders need high level of health care. in every public section, each person produces a lot of data, such that, total number of data about public administration in one nation is very huge. For example, by 2011 there are 3 terabytes of data collected by the US Library of Congress. In 2012, The Obama administration announced that the Big Data research and development initiative. It investigates and addressed that, by using such big data government facing the problems. Six departments were involved for the initiative consists of 84 distinct Big Data programs. In Europe, this situation is repeated .To improve the productivity of governments around the world they are facing unfavouarable circumstances. In public administration, they are more effective. With significant budgetary constraints, in the recent global recession many governments have to provide a higher level of public services. Hence, they would take Big Data as a potential budget resource and develop tools to get alternative solutions to reduce national debt levels and decrease big budget deficits

4.2. Big Data in Business and Commerce

According to the forecasting of [13], for every 1.2 years the volume of worldwide business data across almost companies .For example, in Retail Industry, around 267 million transactions per day in Wal-Mart’s 6000 stores worldwide. Recently, Wal-Mart is collaborated with Hewlett Packard to store 4 peta bytes of data, i.e. 4000 trillion bytes; it is traced from their point-of-sale terminals for every purchase record. With the help of machine learning techniques they have successfully improved the efficiency of their advertising campaigns and pricing strategies. The management of their inventory and Supply chain significantly benefitted from large data warehouse. McKinsey’s Report saying that [15], Big Data functionalities such as higher levels of effectiveness and efficiency, provide the public sector to improve the productivity and reserving the informative patterns and knowledge.

4.3. Big Data In Scientific Research

Many of the scientific areas are already with the development of computer sciences highly data-driven [16]. For example, meteorology, astronomy, social computing [17], computational biology [18] and bioinformatics are based on scientific discovery as massive volume of data is generated with distinct types these science fields.

5. STATE OF THE ART TOOLS AND TECHNIQUES TO HANDLE DATA-INTENSIVE APPLICATIONS

5.1. Big Data Technologies and Techniques

We need to develop new technologies and techniques to analyze the data and to capture the value from big data. Till now scientists have developed various techniques to curate, capture analyze and visualize the Big Data. These technologies and techniques crossed a number of disciplines such as economics, computer science, statistics, mathematics and other expertise. Multidisciplinary methods are required to discover the useful information from Big Data. We will discuss present technologies and techniques to exploit the data intensive applications. To make sense of Big Data, we need tools (platforms). Present tools focusing on three classes, such as stream processing tools, batch processing tools and interactive analysis tools. Many of batch processing tools [20] are based upon the Apache Hadoop infrastructure as follows Dryad and Mahout. For large scale streaming data analytics platforms S4 and Storm are examples. In an interactive environment, the interactive analysis processes the data and allows the users to commit one their own analysis of information. In real time user is connected to PC and he can interact with it. The data can be compared, reviewed and analyzed in graphical format or tabular format or both at the same time. Apache Drill and Google's Dremel are based upon the interactive analysis.

5.1.1. Big Data Techniques

Big data requires outstanding techniques to efficiently process massive volume of data within the limited run times. For example, to explore patterns from their large volume of transaction data, Wal-Mart applies statistical techniques and machine learning. These patterns generate high competing in advertising campaigns and pricing strategies. Taobao ,A Chinese company like eBay, on users' browse data recorded on its website and exploits a good deal of useful information to support their decision-making, it was adopted a massive stream data mining techniques. Big data techniques involved in number of areas such as data mining, statistics, neural networks, machine learning, social network analysis, pattern recognition, signal processing, optimization methods and visualization approaches.

5.2. Statistics

To collect, organize and interpret the data statistics techniques are used. To exploit the casual relationship and correlation ship among distinct objectives. Authors proposed efficient approximate algorithm for large-scale multivariate monotonic regression. It is an approach for estimating functions that are monotonic with respect to input variables. Another trend of data-driven statistical analysis focusing on scale and parallel implementation of statistical algorithms. With the help of statistics numerical descriptions are generated [6]. Statistical learning and Statistical computing are the two hot research sub-fields.

5.3. Optimization Methods

To solve quantitative problems in many areas such as biology, physics, economics and engineering Optimization methods are applied. In [19], various computational strategies are addressed for global optimization problems such as adaptive simulated annealing, simulated annealing genetic algorithm and quantum annealing. Stochastic optimization includes evolutionary programming; genetic programming and particle swarm optimization are useful. Most of the research works are done to scale up large-scale optimization by co-evolutionary algorithms. Real-time optimization is needed in various Big Data application, such as ITSs and WSNs. Parallelization and Data reduction are also alternative approaches in optimization problems.

5.4. Data Mining

Data mining is a collection of techniques to extract useful patterns from data such as Classification and Clustering analysis, association rule mining, and regression, discriminate analysis. It involves the methods from statistics and machine learning. When compared to conventional data mining algorithms Big Data mining is a Challenging issue. Most of the extensions usually relies on analyzing a particular amount of samples of Big Data, and vary in how the sample-based results are used to derive a partition for the overall data. Clustering algorithms such as CLARA (Clustering LARge Applications) algorithm, CLARANS (Clustering Large Applications based upon RANdomized Search), BIRCH (Balanced Iterative Reducing using Cluster Hierarchies) algorithm, etc. To reflect the goodness Genetic algorithms are also applied to clustering as optimization criterion. Social Network Analysis (SNA) is emerged as a key technique in modern sociology, views social relationships in terms of network theory; it consists of nodes and ties. Visualization Approaches are the techniques used to create diagrams, tables, images and other intuitive display ways to understand data. Machine learning is an important subject of artificial intelligence. It is aimed to design algorithms that allow computers to evolve behaviors based on empirical data.

Big Data tools for batch processing

1) Karmasphere Studio and Analyst
2) Jasper soft BI
3) Sky tree Server
4) Pentaho Business Analytics
5) Apache Mahout
6) Tableau
7) Talend Open Studio

8) Apache Hadoop and map/reduce
9) Dryad

Big Data tools for stream processing

1) Storm
2) S4
3) SQLstream s-Server
4) Splunk
5) Apache Kafka
6) SAP Hana

6. CONCLUSION

In a Distributed Systems world, big data started to become a major challenge in the late 1990's due to the impact of world-wide web. Database technology (including parallel databases) was considered for the task, but was found to be neither well-suited nor cost-effective for those purposes. The necessity to process massive quantities of data has never been greater. Not only terabyte and petabyte scale datasets rapidly becoming common place. Gartner defined Big Data as "Big Data is High Velocity, High Volume and High variety information assets require new forms of processing to enable enhanced decision making, process optimization and insight discovery. In the commercial world, business intelligence gathers the data from array of sources. Big Data analysis tools like Map Reduce over Hadoop, HDFS, to assist to organizations better understand their market place and customers hopefully leading to better business decisions and competitive benefits. For engineers building information processing tools and applications, large and heterogeneous datasets which are generating continuous flow of data, lead to more effective algorithms for a wide range of tasks. Web-based applications are encounter big data such as social computing (includes online communities, reputation systems, social network analysis, prediction markets, recommender systems, Internet search indexing. There are various applications of the Big Data such as Big Data in Society Administration, Big Data in Business and Commerce, Big Data in scientific research. Big Data tools for batch processing such as Apache Mahout Tableau, Talend Open Studio and Apache Hadoop and map/reduce and Dryad Big Data tools for stream process such as Splunk, SAP Hana. Big Data Techniques such as Statistical techniques, Optimization Methods and Data mining techniques, Machine Learning Techniques, Classification and Clustering techniques, Regression Analysis techniques, etc were discussed.Algorithm is discussed on Building index for web pages available online and see how Map and Reduce functions can be executed. Input can be considered as a set of documents.

REFERENCES

[1] Puneet Singh Duggal, Sanchita Paul, *"Big Data Analysis: Challenges and Solutions",* International Conference on Cloud, Big Data and Trust, RGPV. November 2013: 13-15: 269-276.

[2] Jefry Dean and Sanjay Ghemwat, MapReduce: A Flexible Data Processing Tool, Communications of the ACM. January 2010: 53(1): 72-77.

[3] Jefry Dean and Sanjay Ghemwat, MapReduce: Simplified data processing on large clusters, Communications of the ACM. 2008: 55: 107–113,

[4] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data", *IEEE Transactions On Knowledge and Data Engineering*. January 2014: 26(1): 97-107.

[5] "IBM What Is Big Data: Bring Big Data to the Enterprise," http://www-01.ibm.com/software/data/bigdata/, IBM, 2012.

[6] C.L. Philip Chen, Chun-Yang Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", Information Sciences, www.elsevier.com/locate/ins, January 2014.

[7] Eric Savitz, Gartner: 10 Critical Tech Trends for the Next FiveYears, October2012 <http://www.forbes.com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five years/>.

[8] Eric Savitz, Gartner: Top 10 Strategic Technology Trends for 2013, October 2012. <http://www.forbes.com/sites/ericsavitz/2012/10/23/gartner-top- 10-strategic-technology-trends-for-2013/>.

[9] Doug Laney, 3d Data management: controlling data volume, velocity and variety, Appl. Delivery Strategies Meta Group (949) (2001).

[10] Paul Zikopoulos, Chris Eaton, Paul. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw Hill Professional, 2011.

[11] Junping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, Cheng Chen, Data-driven intelligent transportation systems: a survey, *IEEE Trans. Intell. Trans. Syst.* 12 (4) (2011) 1624–1639.

[12] Ewaryst Tkacz, Adrian Kapczyn´ ski, Internet: Technical Development and Applications, Springer, 2009.

[13] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, Big data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2012.

[14] Randal E. Bryant, Data Intensive supercomputing: The Case for Disc. Technical Report CMU-CS-07-128, 2007.

[15] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh,Angela Hung Byers, Big data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2012.

[16] Alexander S. Szalay, Extreme data-intensive scientific computing, Comput. Sci. Eng. 2011: 13(6):34–41.

[17] Fei-Yue Wang, Daniel Zeng, Kathleen M. Carley, Wenji Mao, Social computing: from social informatics to social intelligence, *IEEE Intell. Syst*. 2007: 22(2): 79–83.

[18] Jason McDermott, Ram Samudrala, Roger Bumgarner, Kristina. Montgomery, Computational Systems Biology, Humana Press, 2009.

[19] Vikas C. Raykar, Ramani Duraiswami, Balaji Krishnapuram, A fast algorithm for learning a ranking function from large-scale data sets, *IEEE Trans. Pattern Anal. Mach. Intell*. 2008: 30(7): 1158–1170, 200.

[20] Michael Isard, Mihai Budiu, Yuan Yu, Andrew Birrell, Dennis Fetterly, and Dryad: *distributed data- parallel programs from sequential building blocks*, in: EuroSys ’07 Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems. 2007: 41(3):59–72.