IS 601 Management Information Systems

Term Research Paper

On

Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application

Submitted to
Dr. Carlton Crabtree

by

Priyadarshini Arcot
Campus ID: HW51333
MS Information Systems

The Graduate School
University of Maryland, Baltimore County

Table of Contents

1. Introduction

Weather forecasting has always been difficult due to the large number of variables involved and the complicated interactions between those variables. Meteorological forecasters' ability to pinpoint the time and intensity of hurricanes, floods, snowstorms, and other weather disasters has significantly improved due to dramatic improvements in data collection and analysis. When information is accurately collected and analyzed, all organizations in any field, such as oil, money, or manufacturing hardware, generate big data, which may present useful designs to business directors in order to make and develop their companies. The information gathered for evaluation has an impact on data quality because it determines whether a particular method of carrying out an ongoing operation is beneficial in the long run.

The quality of Big Data is extremely relevant and important. We are examining processes such as cleansing to fix as much data as possible. subprocess for Filtering, Integration, and Data Transformation, as well as a noise filter to remove bad data. Big data consistency is really important. We propose that different forms of raw data be evaluated to increase precision in the pre-handling stage, because some bits of data are not used later in the process. These large amounts of data might come from a variety of sources, including business network exchange systems, consumer records, produced by a computer and real-time data sensors from the Internet of Things. All of the information and abilities for their implementation are stored in data in a big data network (Labeeb, K., 2020).

Big Data has emerged as the preeminent method for acquiring, processing, and analyzing data. Quality of Big Data (QBD) has emerged a At the inception phase, the nature of targeted data, such as those arising in social networks and characterized by unstructured data, must be profiled with certain quality dimensions.s a critical factor in ensuring that data quality is maintained throughout all phases of Big Data processing. To support the selection and adaptation of data quality profiles It tracks and registers the effect of every data transformation that occurs during the pre-processing phase in a data provenance repository.
A concept of data quality is required to keep track of data value and relevance, as well as the severity of the impact of the aforementioned preprocessing transformations. This also implies that the quality of a data attribute must be controlled throughout its life cycle because it has a direct impact on the results of the analysis phase (Ikbal, T., 2015).

Big data emphasizes addressing data in the system early on to increase its relevance. The relevance of big data is excellent. Raw data can be abused by a combination of developed, semi-built, and unbuilt data, including degraded qualities that are insufficiently organized. Raw data types are used to improve quality during the preprocessing phase. The process includes cleansing to fix data as much as possible, noise filters to remove bad data, sub integration and filtering, and data transformation. During the acquisition phase, big data is adapted to cost overheads as well as improving accurate data analysis. It enables and prepares businesses to take the next step in their growth strategies (Juneja, G. A., 2019).

2. Literature Review

Big data is a term that refers to massive amounts of structured, semi-structured, and unstructured data that are challenging to process with traditional methods and databases. To turn data into an asset for an organization, high-level programming skills and procedures are used. Machine-generated data and real-time data sensors used in internet of things (IoT) contexts provide this type of information. For the past several years, various research has been adopted in weather monitoring, and while it has improved over time, these finding and implementing tactics have had very little impact on development. Data quality is defined in quality management as the suitability of data for usage or meeting user requirements. Many phase processes will be altered, either positively or negatively, if quality is associated with data at its origin. To increase the quality of data, it is exposed to auditing, profiling, and the application of quality norms. Proper administration of these data quality procedures is necessary to overcome the numerous obstacles that arise when working with large data sets. Big data necessitates well-defined lightweight measuring approaches that can operate in parallel with each phase. Data quality management, monitoring, and control are among these procedures, with the primary purpose of tracking any changes that improve or decrease data quality. Processes for data cleaning, data reduction, data deletion, and various more sub-processes are included in the standard framework. Several data quality issues arise when attempting to apply data quality ideas to large data sets. Data cleansing is used in this procedure to find and correct inconsistencies. You can save time and money by using a data quality selection methodology.(Ikbal, T., 2015).

Data cleaning, for example, addressing missing attributes, balancing a dataset, and removing noise or outliers, are all examples of pre-processing data. Predictive Data Mining refers to supervised data mining methods, whereas Descriptive Data Mining is an unsupervised method that focuses on visualizing the dataset's properties. To improve data quality before using data mining techniques to extract meaningful information In large data sets, data mining is the process of searching for hidden, valid, and potentially helpful patterns. It begins with data preprocessing procedures, which includes a full overview of several data cleaning approaches (Tomar, D., 2014).

Data quality can provide a variety of services to a business, and high-quality data can currently increase an organization's chances of achieving top-tier services. Data quality is affected by the size, speed, and format in which it is generated. various raw data features in the pre-processing stage to improve its quality, Cleansing to fix as much data as possible, Noise filters to remove bad data, Integration and Filtering sub-processes, and Data Normalization Filters can be used to eliminate unattended trash data as well as to traverse the network. The framework allows access to large files that may be organized or, more likely, unstructured or semi-structured, and contain an abundance of data sources and data formats.The application of particular domain rules, automatic rules, or user-defined rules derived from observation can help to improve the accuracy of the prediction system's results. Quality of Big Data a variety of offerings for developing a dependable and precise technique In all businesses, data is a valuable resource, and the quality of data is crucial for managers and operating procedures to discover relevant performance issues

(Sidi, F., 2012).

3. Technical details

Data accuracy must be constantly tracked, managed, and tuned in order to be used in the most efficient and effective way possible. Every field has data that is specific to its market or company. The information will also be useful in determining the customer's opinion. Data must be sensitive to the environment and adapted to the needs of consumers. The information will also be useful in determining the customer's opinion (Sidi, F., 2012). Before it is used, consider how to improve the correctness of the data during the early stages of subsequent processing or analysis. The adequacy of data characteristics such as shape, structure, time, and type variations must be approached from both a consistency and a consistency-based perspective

Many factors will be considered in the development of a data quality system, with key factors including corporate domain, data source(s), and constructed / unconstructed data. The higher the source, the higher the quality of the previously analyzed results. The accuracy of the source determines the quality of the data.

3.1 Dimensions of data quality

The measurement can depend upon the data reliability. There is no standard description of consistency. The meaning can change depending on the market domain, framework, or importance. The most common intrinsic measurements are accuracy, reliability, authenticity, completeness, credibility, and completeness.

3.2 Profiling of data

The goal of profiling data is to create a framework. The rule for ensuring effective data quality evaluation, Multiple iterations in an attempt to cleanse and shift from an unconstructed to a more organized state will analyze and transform data quality in a big data network to maintain data consistency.

3.3 Frameworks for quality of data

This framework establishes guidelines for achieving higher-quality results. The standard framework includes processes for data cleaning, data reduction, data deletion, and several other sub-processes.

3.4 Quality of Big Data

Data quality is a developing and growing field of big data. By gathering data from various business structures, various branches, revenue, income, geography, and location criteria, they will be able to expand geographically into more fields. It is critical for every big data application on the market to ensure and transform data into consistency in order to reliably interpret the data and determine trends that lead to the exact or best possible implementation of potential strategies.

3.5 Pre-processing of Big Data

Big data programs, data pre-processing, and data consistency analysis increase and improve data values. Typically, lifecycle data pre-processing includes the following subsections:

(i) The fortification and Integration of data

Data can come from a variety of sources and be assembled in a variety of ways. Both of these sources of data must be uniformly merged so that the data to be included in the Big Data network becomes a single, final source of data.

(ii) Improvement and Enrichment of data

Data is compiled from a variety of sources and then supplemented with additional data from other sources. the variety of data, which comprises organized, quantitative, and unstructured text data in various formats. Data quality must be properly conditioned to its specific area. Data from many supporting sources is combined. Understanding the data's source is crucial and determines its accuracy or reliability level, as the next phase will evaluate the data using this knowledge.

(iii) Transformation of data

Sub-processes in data processing include acquiring or collecting data from various sources, among others. According to regulatory requirements, data will need to be reformatted, compressed, analyzed, or updated.

(iv) Reduction of data

Data reduction is a mechanism for reducing data volume to a non-redundant state. This tries to improve data storage quality while also saving money by removing data that isn't required and retaining only the components that are required for this specific operation.

(v) Discretization of data

This strategy gathers and separates data into intervals in order to make the best use of the current mining methods and processes.

(vi) Cleansing of data

For improving data efficiency it can be done by removing data that makes data less accessible. The data must be handled appropriately because it is unreliable for speedy analysis. Data cleansing is the process of removing irregularities from current data in order to obtain a data collection that is an accurate and unique depiction of the micro world. Manual data purification is nearly impossible due to the large amount of data collected, as it is time-consuming and prone to errors. The data cleansing process is lengthy and involves multiple steps, including defining quality guidelines, finding data errors, and correcting them.

4. Development and Refinement of Big Data

Pre-processing data and deciding whether to format the data as a single Data Stream are essential for business success. The majority of businesses generate a massive amount of data over the course of numerous projects. According to guidelines specified when data quality requirements were approved, the Data Quality Profile focused on a number of quality assessments, including data source, replication, and anomaly detection (Taleb, I., 2015). In the next stage, data quality profiles are validated using data samples, i.e. data quality criteria are evaluated. The DQDs employed in assessing DQ in traditional data management systems are generally appropriate in the measurement of DQ in BD. This analysis comes to the interesting conclusion that, with the correct probabilistic generalizations in place, refinement may be considered to encompass from raw data and an unresolved issue to a scenario.

Most businesses generate a large amount of data over time and across various initiatives. Data can come from a variety of sources, including internal and external systems, and can be in a variety of formats. Samples of data from various resources are used to validate the Data Quality profiles. These stages are part of the Pre-Processing automation, which is iterated recursively until the Data Quality meets the pre-defined standards. To enable user participation in order to evaluate and enhance the data quality profile for each domain on a regular basis.

It is a data warehousing process that extracts data from external sources. Most ETL systems have a graphical interface for creating ETL workflows and automating their execution. appropriate data quality requirements in accordance with a pre-defined or approved 'description of finished' A final gateway procedure will necessitate user experiences that examine and improve the quality of data profiles according to domain on a regular basis in order to attain the desired consistency (Labeeb, K., 2020).

5. Risks and limitations

The risks associated with deploying big data are mostly dependent on the data collection process; if there are any errors or incorrect assumptions made while processing the data, it will have a significant influence on all of its applications. The primary stumbling block is the high expense of adopting modern technologies.

These technologies require a significant investment. Although the former's participation is the most important necessity for big data analysis, their inputs are still required for certain elements. Upgrading the machinery is more expensive, and there is a good risk that the data obtained from these devices will be inaccurate.

It is difficult to reverse the process if there is a technical issue with the data from the machines, which is a waste of resources and time. Although the chances of the aforementioned limitation and risk occurring are extremely remote, they cannot be ruled out entirely.

6. Suggested course of action

The concept of a big data framework in weather forecasting has gotten a lot of attention in recent years, and the accompanying paper recommends the best methods for obtaining accurate findings and using data quality to get and process correct forecasts in weather forecasting. The data assessment approach aids in the domain-by-domain examination of data quality. The forecast grows more accurate as more data is processed. Big Data is a concept that describes the data that can be acquired and used to benefit any business (Bansal, S. K., 2014).

7. Conclusion

This article gives an outline of how big data is changing the industry and what are the best ways that can be used to incorporate multiple technologies to gain better results. Big data applications can help industries to leverage their business with the help of data acquired over some time. The evaluated study was based on recent global warming concerns and attempted to design a weather forecasting system using Big Data collected from various sources. It emphasizes the importance of addressing data in a Big Data system early on in order to maximize its relevance.

Technologies like IoT, cloud computing are great sources for getting live data. With the help of these technologies, we can forecast the weather. The information gathered to be analyzed affects data quality because that data determines whether a specific method of conducting the ongoing process is useful or not in the long run. In the pre-handling stage, raw data should be analyzed to improve precision. To improve information consistency we break down and model huge data in order to reduce overhead costs and build a strong grasp of results. (Labeeb, K., 2020).

The accuracy, integrity, formatting, and application of rules and information are all strengthened by the weather forecasting domain and the special domain rules, automatic rules, or user-defined rules learned through observation, aids in improving the accuracy of forecast system results. The data assessment process aids in determining the quality of data based on the domain. The higher the quality of the pre-processed data, the better the analyses or forecasts. It has the potential to enable and prepare enterprises to take significant steps forward in their growth and future objectives.

8. Annotated Bibliography

Labeeb, K., Chowdhury, B. Q., R. B. Riha, R. B., Abedin, M. Z., Yesmin, S., and Khan, M.

N. R., "Pre-Processing Data In Weather Monitoring Application By Using Big Data Quality Framework," ***2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*****, 2020, pp. 284-287, doi: 10.1109/WIECON-ECE52138.2020.9397990.**

The paper describes the evolution of Big data gathered from multiple sources to design a system which is capable of forecasting the weather conditions based on global warming. The author describes the study of weather applications and the attempt to make use of Big data where the author developed a tracker for weather forecasting to explain data preprocessing processes. It makes use of IoT sensors for different weather variables, remote APIs, micro-services for mobile parameters, position input information for travel directions, etc. The app consists of evidence of change of temperature exhibited by global warming which will track the fluctuations above 1.5 degree centigrade and activate warning alarms with the use of these several small locations in a wider area, and better forecasts can be generated correctly.

In this application we can only know how much data is collected and processed from sources, devices and sensors to track temperature. It is also critical for the data with a Big data lifecycle to use preprocessing data. Filters can also be used to eliminate unattended waste data, and the framework exposes large files that can be organized. The application of special domain rules, automatic rules, or user-defined rules learned through observation, aids in improving the accuracy of forecast system results. The higher the quality of the pre-processed data, the better the analyzes or forecasts produced by the device. This paper has all tasks, techniques and models for a dataset which will be useful for my research work.

Taleb, I., Dssouli, R. and Serhani, M. A., "Big Data Preprocessing: A Quality Framework", ***2015 IEEE International Congress on Big Data (BigDataCongress)*****, 2015.**

In this paper Big data preprocessing quality framework attempts to apply a large data set to address large quality concerts. The author describes the quality of Big data(QBD) at the preprocessing phase which includes sub-processes like cleaning,integrating,filtering and normalization. The model is used to support data quality profile selection and adaptation. It also tracks and registers every data transformation preprocessing phase which can be evaluated by the EEG data set. The data quality selection model evaluation can be done by a large EEG dataset. The author explains that when new rules are applied or deleted the DQP on data sample sets faster data quality. The drawback for this component is it lacks quality rule diversity and Prior to data analysis, these rules are used as pre-processing tasks. For the pre-processing of an EEG dataset, we examined the data quality profile selection. It will be a value added feature by having DQP for DQR with a mechanism to insert,rate and populate.

In this the author aligns that the DQP is generated on data samples rather than entire data sets, allowing for faster data quality review and immediate updates when new quality rules are added. A value-added feature will be having a DQP repository for data quality rules with tools to populate, update, query, and rate these rules automatically. The results underline the importance of addressing QBD early in the Big Data processing process. This paper will be useful to understand the quality framework and its techniques.

Juneja, G. A. and Das, N. N., "Big Data Quality Framework: Preprocessing Data in Weather

Monitoring Application", ***2019 International Conference on Machine Learning Big Data Cloud and Parallel Computing (COMITCon)*****, pp. 559-563, 2019.**

This paper proposes a preprocessing framework to address quality of data in weather monitoring and forecasting by taking global warming into consideration and raising notification to warm users in advance. It emphasizes addressing data in the Big data system to magnify its relevance at an early stage. Quality of big data is great in relevance. The author explains that various raw data is used to improve quality in the preprocessing phase. The process consists of cleansing to fix data as feasible, noise filters to remove bad data as well as sub integration and filtering along with data transformation. The Big data during acquisition phase is adapted to cost overheads and also by improving accurate data analysis. It enables and prepares organizations to take leap forward with their growth strategies.

The author describes the Pre-Processing Framework to handle data quality in a weather monitoring and forecasting application. Data aids in the acquisition, processing, and analysis of vast amounts of heterogeneous data in order to produce useful outcomes. a QBD model with methods to aid in the selection and adaptation of data quality profiles Addressing data in the early stages of a Big Data system to increase its relevance. This can indeed enable and prepare firms to take a big step ahead in terms of future plans.

Sidi, F., Shariat Panahi, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H. and Mustapha, A., "Data quality: A survey of data quality dimensions", ***2012 International Conference on Information Retrieval Knowledge Management (CAMP)*****, pp. 300-304, 2012.**

This study focuses on data quality dimensions and proposes a framework for measuring dimensions and improving process quality using a proposed framework combining data mining and statistical techniques. According to the author, data is a valuable resource in businesses, and it is vital to link data to connected concerns. Data may be improved by recognizing various features of data for dimensions, types, and approaches. Existing surveys from 1985 to 2009 revealed the data quality. The goal of this survey is to identify a link between data quality and framework. To increase the quality, we need to assess the interdependence of the data quality aspects described. The purpose is to obtain an agreement on matters on which everyone agrees. The meaning can alter depending on the market domain, context, or significance.

The author explains that the goal is to reach a consensus on the issues that everyone agrees on. Depending on the market domain, context, or significance, the meaning can change. Data quality metrics are used to determine data reliability.

Cichy, C., Rass, S., "An Overview of Data Quality Frameworks", ***Access IEEE*****, vol. 7, pp. 24634-24648, 2019.**

This article compares data quality frameworks used in a variety of business environments in terms of definition and improvement of data quality, as well as providing a decision guide to data quality frameworks. There was some variety in data quality definitions because the frameworks all chose various data quality aspects to be relevant. The majority of frameworks are built to deal with structured and semi-structured data, with only a few exceptions dealing with unstructured data. In the same way that a decision tree narrows down the choice to a suitable framework from a particular situation, this paper assists in identifying frameworks that are suitable and the number of crucial components. The disadvantages include a lack of thorough predictions of the effects of poor data quality, as well as the lack of consequences and interactions of data quality with a sophisticated statistical base.

In this paper we discuss various weather prediction models proposed by different researchers. It makes use of water resource and rainfall prediction which is important for statistical and predictive models for rainfall forecasting that is available and weather damage prediction is an important issue where they can forecast the particular location. It provides less accuracy on a large scale due to climate. This paper will help me with my research work by understanding the data quality framework and its implementation.

Bansal, S. K., "Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration", ***2014 IEEE International Congress on Big Data (BigData Congress)*****, pp. 522-529, 2014.**

This paper proposes a semantic Extract-Transform-Load (ETL) architecture for integrating and publishing open linked data from many sources. Because we need technology and tools to find, transform, analyze, and visualize data before it can be used for decision-making, the majority of this data remains inaccessible to people. The majority of this data is inaccessible to users because we require technology and tools to find, transform, analyze, and visualize data before it can be used for decision-making. the creation of a distributed Web of data utilizing the Resource Description Framework (RDF) as the graph data model and SPARQL as the semantic query language to extract meaningful knowledge and information from the combined data In the industry, the Extract-Transform-Load (ETL) process has been used for data integration. The development of a semantic data model to serve as a foundation for the integration and comprehension of knowledge from multiple sources. This paper helps me understand the framework for big data integ Search/query engines and analytic tools for Big Data must be used effectively and creatively to develop smart and sustainable ecosystems.ration and techniques.

The author briefs the details with the study presents a semantic Extract-Transform-Load (ETL) architecture for integrating and publishing data from diverse sources as open linked data using semantic technologies.This paper will help me with my research work by understanding the data quality framework and its implementation.

Tomar, D., & Agarwal, S. (2014). A survey on pre-processing and post-processing techniques in data mining. ***International Journal of Database Theory and Application*****,** ***7*****(4), 99-128.**

This paper gives us a brief survey on pre-processing and post-processing techniques. It begins with pre-processing techniques, such as a detailed description of various data cleaning approaches, imbalanced data handling, and dimensionality reduction. In order to improve the quality of these data, pre-processing techniques must be used. Several visualization techniques, including scatter plots, parallel coordinates, and pixel oriented techniques, are discussed in this paper. The paper also includes detailed descriptions of three visualization tools, DBMiner, Spotfire, and WinViz, as well as a comparative evaluation based on specific criteria. Knowledge Discovery in Databases (KDD) refers to various processes for extracting useful information from large amounts of data. It also emphasizes the research opportunities and challenges associated with the Knowledge

Discovery process.

In this paper the author explains about the Several visualization techniques that are discussed in this study, including scatter plots, parallel coordinates, and pixel oriented techniques. The study also offers detailed explanations of three visualization tools, DBMiner, Spotfire, and WinViz, as well as a comparison of their performance based on a set of criteria.

Glowalla, P., Balazy, P., Basten D., and Sunyaev, A., "Process-Driven Data Quality Management – An Application of the Combined Conceptual Life Cycle Model", ***2014 47th Hawaii International Conference on System Sciences (HICSS)*****, pp. 4700-4709, 2014.**

This paper describes Process-driven data quality management, which enables long-term data quality improvements both within and outside the IS domain, and is becoming increasingly important. Existing process modeling approaches do not explicitly link data quality dimensions to the production of context-specific information products (IP). As a result, we offer a process-driven implementation of the combined conceptual life cycle (CCLC) model for process exploration and data quality improvement. a detailed case study of a medium-sized company that launched a process optimization initiative to improve data quality The results demonstrate the approach's benefits and limitations, allowing practitioners to tailor the approach to their specific needs.

The author examines BD and the most often used DQDs for BD, which serve as a foundation for assessing and evaluating the quality of BD. The most frequent DQDs used for BD are Accuracy, Consistency, Completeness, and Timeliness, according to the article. There's still a lot to learn about BD for DQDs in order to test BD quality effectively and efficiently.

Hu, H., Wen, Y., Chua, T. -S., and Li, X., "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", ***IEEE Access*****, vol. 2, pp. 652-687, 2014.**

This paper presents a literature review and system tutorial for big data analytics platforms, with the purpose of offering a broad overview for non-expert readers and inspiring advanced audiences to design their own big-data solutions. The four steps of the big data value chain include data generation, data capture, data storage, and data analysis. During the big data acquisition phase, typical data gathering technologies were studied, followed by big data transport and big data pre-processing approaches. a number of cutting-edge and representative computation models It examines a variety of methodologies and mechanisms from the scientific and industry worlds in depth. During the data analytics phase, several data analytics approaches are organized by data attributes.

The author explains about the method for breaking down large data systems into four distinct modules: data generation, data acquisition, data storage, and data analytics.This paper consists of models, techniques of Big data and its phases.

Kotturu, P. K., and Kumar, A., "Data Mining Visualization with the Impact of Nature Inspired Algorithms in Big Data," ***2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*****, 2020, pp. 664-668, doi: 10.1109/ICOEI48184.2020.9142979.**

This paper explores the possibilities and elaborates on the background in the direction of big data exploration with the impact of various computational methods.To examine computational methods in the context of machine learning and data mining. Scaling and analyzing performance parameters are required. Data mining hybridization in terms of cloud computing association. It offers detailed analysis and visualization in the direction of various aspects of data mining in the context of Big Data. The methodological advancement, as well as the problem statements, have been analyzed based on these aspects. The incorporation of an approach aimed at utilizing aspects of cloud analytics and big data for data access and synchronization mechanisms To investigate the possibilities of cloud computing and big data using various parameter variations and validations. On the basis of these considerations, the methodological progress as well as the problem statements have been examined. This will aid in the exploration of new insights in this domain.

The author describes the comprehensive generalization of empirical and methodological features, demonstrating the integration of the methodology in the direction of using cloud analytics and big data for data access and synchronization mechanisms.

Jain, H., and Jain, R., "Big data in weather forecasting: Applications and challenges," ***2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*****, 2017, pp. 138-142, doi: 10.1109/ICBDACI.2017.8070824.**

This paper gives a brief survey on the requirement to analyze a vast set of data. Big data in weather forecasting will give several benefits, including saving lives, improving quality of life, and decreasing risks. The numerical weather prediction model (NWP) and various permutations of different models are used to predict weather. Industries have improved dramatically in recent years as a result of the introduction of weather forecasting in them. Data volume and variety are rapidly increasing, posing a significant challenge in weather forecasting, as it is now difficult to combine these data to provide accurate forecasts. Since we have such a large amount of complicated data, data transportation, storage, and administration are becoming an issue, as well as increasing overheads.

In this paper the author briefs about the ideal solutions to these difficulties and have more efficient and reliable applications that can save lives, improve quality of life and business, minimize risks, and increase profitability if we can grasp the nature of these applications and challenges. Scaling and analysis of performance metrics are required.

Haupt, S. E.,and Kosovic, B., "Big Data and Machine Learning for Applied Weather Forecasts: Forecasting Solar Power for Utility Operations," ***2015 IEEE Symposium Series on Computational Intelligence*****, 2015, pp. 496-501, doi: 10.1109/SSCI.2015.79.**

The prospects for forecasting wind and solar resources are explored in this research. Accurate forecasting of these meteorological variables is a big data problem that necessitates a plethora of disparate data, multiple models, each applicable to a specific time frame, and the application of computational intelligence techniques to successfully blend model and observational data in real-time and deliver it to utilities and grid operators.

These applications are shifting to cloud computing frameworks, which introduces new complications, according to the author. The difficulties of separate data arriving at different times and necessitating blending to generate real-time projections will grow much more complicated as we seek to deploy on a larger range of architectures. As we look to integrate more wind and solar energy into the grid, better forecasts of renewable energy factors are becoming increasingly crucial. As the need for these forecasts increases, so will the number of solutions available to solve.

Ismail, K. A., Abdul Majid, M., Mohamed Zain, J. and Abu Bakar, N. A., "Big Data prediction framework for weather Temperature based on MapReduce algorithm," ***2016 IEEE Conference on Open Systems (ICOS)*****, 2016, pp. 13-17, doi: 10.1109/ICOS.2016.7881981.**

In this paper the study and knowledge of how weather temperature changes over time in a particular location or country can be useful for a variety of reasons. Data processing is time consuming and complex due to the volume and velocity of data in each sensor. This project intends to create a MapReduce-based analytical Big Data prediction system for weather temperature. MapReduce is a framework for employing a large number of commodity computers to run highly parallelizable and distributable algorithms across enormous datasets. To analyze sensor data, often known as Big Data, MapReduce with Hadoop is an effective solution. By combining Hadoop with MapReduce, the scalability issue is eliminated. Data processing speeds up as more systems are added to the distributed network.

The author aligns that MapReduce and Hadoop will continue to improve and expand their capabilities. The use of these technologies to large-scale data analysis has the potential to improve weather forecasting significantly.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. ***Information sciences*****,** ***275*****, 314-347. https://doi.org/10.1016/j.ins.2014.01.015**

The paper provides a brief introduction of how Big Data is critical for producing economic productivity and evolutionary advancements in scientific disciplines, allowing us to make considerable progress in a range of fields. Big Data also comes with it a bevy of concerns, including data collecting, storage, processing, and visualization, as well as current approaches and technology, as well as possibilities and challenges. We also go over some possible solutions, such as cloud computing and quantum computing. Without a doubt, future corporate productivity and technological competition will unavoidably lead to Big Data research. Big Data is extremely beneficial for enhancing business efficiency and making evolutionary gains in scientific domains.

The author of this paper seeks to present a comprehensive overview of Big Data, including Big Data applications, Big Data potential and challenges, and current state-of-the-art approaches and technology for dealing with Big Data.

Tang, N. (2014, September). Big data cleaning. In *Asia-Pacific Web Conference* (pp. 13-24). Springer, Cham. https://doi.org/10.1016/j.ins.2014.01.015

In this paper the author explains that Big Data issues affect a wide range of fields and sectors, from economic and business activity to governmental administration, from national security to scientific study in a variety of fields. Big Data also brings with it a slew of issues, including issues with data acquisition, storage, processing, and visualization. The ultimate goals are to foster Big Data science research and innovation. A new wave of scientific revolution is set to commence with Big Data, which is the next frontier for innovation, competition, and productivity.

The author aims to provide a detailed look at Big Data, covering Big Data applications, Big Data potential and difficulties, and the current state-of-the-art methodologies and technologies used to address Big Data issues and provide a high-level review of Big Data issues, covering Big Data potential and challenges, as well as current approaches and technologies and there are various potential solutions to the problem, including cloud computing and quantum computing.