Knowledge-Based Systems 000 (2017) 1–17

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Towards data analysis for weather cloud computing

Victor Chang*

IBSS, Xi'an Jiaotong Liverpool University, Suzhou, China

ARTICLE INFO

Article history:
Received 21 October 2016
Revised 24 February 2017
Accepted 1 March 2017
Available online xxx

Index Terms:
System and application for weather computation
Temperature forecasting and distribution
Mapreduce
Weather data visualization
Polar vortex
Weather data science

ABSTRACT

This paper demonstrates an innovative data analysis for weather using Cloud Computing, integrating both system and application Data Science services to investigate extreme weather events. Identifying five existing projects with ongoing challenges, our aim is to process, analyze and visualize collected data, study implications and report meaningful findings. We demonstrate the use of Cloud Computing technologies, MapReduce and optimization techniques to simulate temperature distributions and analyze weather data. Two major cases are presented. The first case is focused on forecasting temperatures based on studying trends from the historical data of Sydney, Singapore and London to compare the historical and forecasted temperatures. The second case is to use five-step MapReduce for numerical data analysis and eight-step process for visualization, which is used to analyze and visualize temperature distributions in the United States, before, during and after the time of experiencing polar vortex, as well as in the United Kingdom during and after the flood. Optimization was used in experiments involved up to 100 nodes between Cloud and non-Cloud and compared performance with and without optimization. There was an improvement in performance between 20% and 30% under 60 nodes in Cloud. Results, discussion and comparison were presented. We justify our research contributions and explain thoroughly in the paper how the three goals can be met: (1) forecasting temperatures of three cities based on evaluating the trends from the historical data; (2) using five-step MapReduce to achieve shorter execution time on Cloud and (3) using eight-step MapReduce with optimization to achieve data visualization for temperature distributions on US and UK maps.

1. Introduction

Data Science is an interdisciplinary area that allows experts in different domains to study and work together [1-2]. Outputs of all different types of work are in the form of the data, which comes in different types of formats. This allows experts in different disciplines to investigate the meaning of data which can be from the same or different disciplines. There are five common characteristics for Data Science: volume, velocity, variety, veracity and value. Volume refers to the size and quantity of the data that have been processed and stored. Velocity is the rate in which the data has been created, processed and analyzed. Variety refers to the different types and formats of data available and ready for further analysis. Veracity is the accuracy and validity of the data analysis. Value is the added value offered by the Data Science, such as allowing organizations to stay competitive and efficient [1–3]. It has become apparently obvious that the processing, analysis and presentation of data outputs will be essential to a growing number of sectors involved.

In order to analyze data successfully, the infrastructure, platform and software as a service approach should be adopted. With regard to infrastructure and platform, system dependency for Data Science allows the developers to design and build systems and understand the relationship between different clusters, jobs, nodes and virtual machines with the aim to work out the best possible recommendations for different scenarios [3,4]. For software as a service, software dependency for Data Science allows the developers to design and develop software and understand the relationship between libraries, algorithms, APIs, commands, outputs and user interfaces [5–7]. For successful development, the dependency on both system and software are essential to ensure that services can meet expected demands and deliver required tasks. The MapReduce framework can be used as an example. A function called map() can be written that can send jobs to different nodes and return results back to the system, whereby the intelligent software can use reduce() function to map to the related groups together and produce outputs based on the collaboration of map() and reduce() functions. The collaborative system and software dependency can achieve development of different

* Correspondence to. .Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
E-mail addresses: ic.victor.chang@gmail.com, Victor.Chang@xjtlu.edu.cn, vchang1_76@yahoo.co.uk

http://dx.doi.org/10.1016/j.knosys.2017.03.003
0950-7051/

types of services known as "Everything as a Service" (EaaS). The concept of EaaS is based on the development and integration of Infrastructure, Platform and Software as a Service to ensure the joint delivery of all the system and software requirements. For example, Business Intelligence as a Service (BIaaS) is developed based on the pipeline-method architecture that allows the output of the first service becomes the input of the second service [8].

Weather as a service is a good example that need strong collaboration between system and software dependency for Data Science. Often such services require supercomputers, intelligent algorithms, data services, visualization technique to jointly deliver. Weather computing will require an excellent dependency between systems, between different parts of the software and between system and software to ensure that results are fast, accurate, responsive and interactive [9,10]. Vigorous scientific processes aided by the use of advanced technology help meteorologists to make weather predictions. Weather information is particularly useful for the general public to make relevant plans, such as going for trips in the sunny and bright weathers and avoiding trips in snowy and flood conditions. In the event of extreme weathers, this information is vital to the general public. For example, the UK experienced the wettest winter flood in 250 years between December 2013 and February 2014 [11]. Thousands of the residents could evacuate to suitable places in advance to avoid destructive impacts by flood, which damaged houses, towns and buildings in various parts of the UK. The US experienced one of the coldest winters due to the impact of the polar vortex in the same period as the UK, causing several states to experience sub-temperatures below −20 C and −30 C. In the similar period, Southern China experienced one of the warmest winters with the mean temperature of 21 C [12]. Some scientists suggest that the extreme weather conditions will become a norm rather than a possibility of less than 1% due to the global warming and intense human activities [13–15]. What are the impacts of the unpredictable weathers? Unpredictable weathers have become more common in the past few years in the UK. These included the driest summer in 2006 and 2013, big freeze in the UK in 2009 and 2010 winters and the wettest summer in 2011. Some months have "abnormal" temperatures such as the warmest April in 2013 (average of 16.7 C) and the potentially the second coolest summer in 2011 (13.6 C).

The first group of scientists using computational weather forecasting was John von Neumann and his colleagues, who undertook the first experimental weather forecast on the ENIAC computer in the late 1940 s [10]. Within a decade of their work, numerical models became the foundations of weather science and also a discipline in computer science. According to [14,16], each year the US had an economic loss averaging more than US$13 billions due to the extreme weathers. This estimated figure is likely to be up due to the increased frequency of extreme weather events. The improved ways of conducting weather modeling and forecasting are essential to the development of weather science. The e-Science community has demonstrated weather applications. However, thousands of CPUs and expensive infrastructure have to be deployed [17]. Post statistical processing methods such as

The use of Cloud Computing can reduce costs and the scale of deployment while being able to compute weather simulations. We will illustrate an example how to achieve performance and visualization to analyze weather data and conduct performance evaluation. Pioneering methods such as Cloud Computing, Big Data Analytics and dependency between Cloud and Big Data should be investigated to make weather science affordable, accessible and technically viable.

Our paper presents a case on how to use Cloud Computing to process the data and Big Data Analytics to present the results based on the integrated dependency approach. The breakdown of our paper is as follows. Section 2 presents the related models for weather forecasting and the background theories supporting these models. Section 3 illustrates the architecture, system design and deployment to perform weather computing. Section 4 shows the results of weather computing with three case studies and compare performance between Cloud and non-Cloud. Section 5 demonstrates the data visualization dealing with the extreme weathers in the US and UK. Section 6 compares our work with similar approaches and summarizes our contributions. Section 7 presents Conclusion.

2. The background theories

This section describes the background theory associated with weather science. Related work is described as follows. First, Campbell and Diebold present their weather forecasting model [18]. They define the term "weather derivatives" and apply the concept to financial derivatives. This includes the use of volatility to illustrate the concept of time-series waves, which corresponds to the daily average temperature between 1996 and 2001 in four cities. Temperatures can be modeled as a normal distribution curve to show the likelihood of the temperatures. Temperature forecasting can adopt financial forecasting method to estimate temperatures in Atlanta, Chicago, Las Vegas and Philadelphia. Results confirm that the most of expected temperatures is within 95% confidence interval of the actual temperatures. Second, Plale et al. [19] explain how two major weather forecasting systems, CASA (Collaborative Adaptive Sensing of the Atmosphere) and LEAD (Linked Environment for Atmospheric Discovery) can interact with each other's data in real time. They present the system architecture between each system and core technologies. They explain how LEAD can use workflow applications to derive their forecast, even though they do not show technical details. They state the use of XML, Meteorological command and control and Blackboard can achieve interactions.

Third, Li [20] describe the pipeline method that utilizes the e-Science principles and use Hadoop algorithm to process applications and data in their public cloud. This is a pioneering approach for Cloud Computing with results supported their method. However, their method has not been applied to other e-Science related specialization such as Weather Science. Fourth, Droegemeier et al. [12] demonstrate the use of Service-Oriented Grid Computing to enable dynamic weather computation. The architecture they use is a LEAD system. They demonstrate the case of radar reflectivity with several examples about Arkansas. These include the precipitation intensity at the eleven, nine and five hours of forecast. They include the precipitation forecast on January 29, 1999 and March 29, 2000. They also explain their workflows approach towards the ingesting and analysis of their data. Fifth, Demirkan and Delen [21] use their service-oriented decision support system to analyze data, including geospatial data, with the analytics and big data in the cloud approach. They have explained the architecture and their high-level method, without revealing much of their implementation process and performance evaluation. Sixth, Gao et al [22] explain more detailed geospatial data approach by using Hapdoop to process big geo-data. They explain their architecture and two proposed algorithms, which are their main contributions to analyze big geo-data. They present all datapoints and map them on the USA map to show their outputs. They have only done two performance tests. Last, proposal by Baldauf et al [23] can perform post statistical processing for weather forecasting and services.

However, these six existing projects demonstrate current challenges to be resolved. First, Campbell and Diebold [18] do not explain enough theories about how the financial derivatives can be applied to temperature forecasting. There is a need to consolidate. Second, the work in [19] is not an interaction but information exchange. There is a need to use any services to process data and analyze the results. Third, both CASA and LEAD are expensive

Please cite this article as: V. Chang, Towards data analysis for weather cloud computing, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.03.003