Automated Data Analysis Tools for Whistler Blackcomb

Introduction

In 2009, Whistler Blackcomb decided to accurately measure and monitor their power consumption, move towards a net zero operating footprint, and combat climate change [1]. As of 2019, Whistler Blackcomb has real-time energy monitoring systems, and has backlog of incoming data to be processed. Whistler Blackcomb has a team of data technicians, and cannot afford hiring more technicians. Local data servers are reaching capacity every 6 months, and adding new servers have increased electricity costs, and reduced floorspace needed for other services. The budget for the data technician payroll, floorspace, electrical and server costs have been exceeded.

To reduce payroll, electricity, and server costs, and reduce floorspace, data must be moved to remote servers, and automated data analysis systems need to be implemented. To create automated data systems that meet company requirements, 3 software libraries called 'Pandas,' 'Spark,' and 'Dask' have been critiqued, and Spark has been determined the software library of choice.

Data Analysis Software Library Comparison

Data Programing Tools	Pandas	Spark	Dask
Size of datasets	Megabytes/Gigabytes	Gigabytes to terrabytes	Gigabytes to terrabytes
Well documented	Yes	Yes	No
Conciders Parellelization of Cores	No	Yes	Yes
Administrative Costs	Free	Relitive to data size	Relative to data size
Optimized for SQL Databases	No	Yes	No
efficient for queries under 100ms?	Yes	No	No
Able to Output data in graphs	Yes	Yes	No
Able to remotely access data	Yes	Yes	Yes

Problem Description

Upper management requirements for the data analysis software is as follows:

- Software must be able to access remote SQL servers
- Throughput must be fast enough to handle complex queries
- Software must use Python language (known by most staff)
- Costs must be lower than what is already implemented
- Software must be able to output visual graphs

Definitions

Dataframe: a table-like data structure that holds data in memory [2]. A dataframe is similar to an excel spreadsheet.

Data Monitoring Systems: heating, lighting, other electricity outputs, and fuel consumption for snowmobiles and mountain grooming machines are monitored with sensors. Sensors stream relevant data in real-time to our servers.

Parallelization/Concurrency: In the scope of data analysis/manipulation, parallelization allows data to be processed in parallel, ie. across several computer cores at once, rather than a single core [3]. Parallelization increases efficiency by distributing work, similar to hiring more employees to work more efficiently rather than having one employee work harder.

SQL Databases: Structured Query Language Databases. An efficient and most used querying language used on databases.

Throughput: The amount of material passing through a system or process. Throughput is a good indicator of speed in a system.

Evaluation

Pandas

Pandas data manipulation and analysis library is written in the Python language using 'C' data structures to increase throughput [4]. Data is processed in 'dataframes' used for machine learning [4]. Pandas runs on a single core and does not use parallelization.

Spark

Spark is a data manipulation engine written in Scala with dataframe support for machine learning and SQL databases [5]. Spark enables parallelization of cores, and streams of data can be concurrently accessed from different storages locations. Python, Java, Scala, and R languages are all supported by Spark [6].

Dask

Dask is the newest data manipulation library. Like Spark, Dask allows for parallelization. Dask is built upon pandas dataframes, and allows for complex computations. Dask does not support SQL databases natively, and is not well documented [7].

Evaluation of Requirements

Software must be able to access remote SQL servers: All software libraries contain protocols capable of connecting to remote servers, however Spark is the only library that contains native functions for making queries to SQL databases [8]. Functions can be written to connect to SQL databases in Pandas and Dask, but costs will be higher than using Spark functions natively. Spark is the best library for this requirement.

Throughput must be fast enough to handle complex queries: Panda's data structures are 'C' based, meaning data processes will run quickly compared to Spark and Dask. If too much data is processed at one time, local computer memory may fail. Spark and Dask can use parallelization to distribute the work to different cores, and will increase throughput. Whistler Blackcomb will process too much data for Pandas to support, so Spark or Dask are the best libraries for this requirement.

Software must use Python language: All three libraries have support for Python. The chosen software library will need to be well documented; good documentation will reduce complexity when transitioning to the new systems. Dask is the newest software library and does not have well written documentation yet. Pandas and Spark have great supporting documentation and will be the best recommendation for this requirement.

Costs must be lower than what is already implemented: Implementing Pandas on a local server will be free. Since data is moving to remote servers, many cloud based services support Spark and Dask. Moving data to remote servers will cost monthly administration fees, but removing local servers would allow new services to move in place, and generate a higher income. New service revenue, such as from a new Cafe, will pay for the cost of remote servers. Spark and Dask are the best libraries to support this requirement.

Software must be able to output visual graphs: Pandas and Spark can output graphs natively [9]. Dask will need functions to be written, and increase costs. Pandas or Spark will best support this requirement.

Conclusion

The Spark data analysis software library is the recommended choice as it meets all of the proposed requirements, and in all cases is the best choices. Spark can access remote SQL servers natively without need to create functionality; Spark supports panelization to distribute data processing tasks across many servers; Spark is supported by Python and has well written documentation; graphs can be produced for management. Moving data to remote servers supported by Spark will incur administrative costs per month, but removing local servers will allow a Cafe to be built which will increase revenue. It is recommended that the data technicians be removed with severance, and the new data analysis system will require 2 IT technicians in their place.

References

[1] Whistler Blackcomb Holdings, "Our Mission: Environment" [Online]. Available: https://www.whistlerblackcomb.com/explore-the-resort/about-the-resort/environment.aspx [Accessed Nov 16, 2019].

[2] Oilshell, "What Is a Data Frame? (In Python, R, and SQL)" [Online]. Available: https://www.oilshell.org/blog/2018/11/30.html [Accessed Nov 16, 2019].

[3] Computer Hope, "Parallelization" [Online]. Available: https://www.computerhope.com/jargon/p/parallelization.htm [Accessed Nov 16, 2019].

[4] Wikipedia Foundation, "pandas (software)" [Online]. Available: https://en.wikipedia.org/wiki/Pandas_(software) [Accessed Nov 16, 2019].

[5] Anaconda Inc, "Comparison to Spark" [Online]. Available: https://docs.dask.org/en/latest/spark.html [Accessed Nov 16, 2019].

[6] MapR Technologies Inc, "What Is Apache Spark?" [Online]. Available: https://mapr.com/blog/spark-101-what-it-what-it-does-and-why-it-matters/ [Accessed Nov 16, 2019].

[7] Anaconda Inc, "High level performance of Pandas, Dask, Spark, and Arrow" [Online]. Available: http://matthewrocklin.com/blog/work/2018/08/28/dataframe-performance-high-level [Accessed Nov 16, 2019].

[8] "Integration with Cloud Infrastructures" [Online]. Available: https://spark.apache.org/docs/latest/cloud-integration.html [Accessed Nov 16, 2019].

[9] "GraphX" [Online]. Available: https://spark.apache.org/graphx/ [Accessed Nov 16, 2019].