Weather Warning System by Stream Processing
Rainfall and Wind Speed Sensor Data

Mustafa Sert, Aşkın Bilgi, Dr. Özgür Yılmaz
msert@mgm.gov.tr, abilgi@mgm.gov.tr, ozgur@dilişim.com

Abstract
In this experimental work, a prototype weather-warning system was implemented to process the streaming sensor data using big data technology mainly Spark and produce weather alarm levels according to the meteorological specifications that would support Turkish National Weather Service's goal. The scope of the work was to demonstrate the tools and techniques to perform stream processing on rainfall and wind speed sensor data that flows from a distributed Automatic Weather Observation System (AWS) network around Turkey. We propose a weather-warning system algorithm that uses Spark which is a fast and general processing engine compatible with Hadoop. Spark is designed to perform both batch processing and new workloads like streaming, interactive queries and machine learning. This big data technology provides highly scalable and fault-tolerant streaming processing while providing near real time processing.

The scope of the work was to demonstrate the tools and techniques to perform stream processing on rainfall and wind speed sensor data of TSMS and produce weather alarm levels accordingly. The exercise also aimed to integrate the dataset and the result set with the data discovery tool in order to present the findings of the analysis.

Keywords: big data; Spark; distributed computing; storm analysis; rainfall; wind

Introduction
The volume, the velocity and the variety of data have been increasing continually in the last decade. Not only capturing and storing the streaming data is a challenge but analyzing and visualizing the streaming data is also a big challenge for organizations. As the value can quickly degrade the value in the data needs to be captured in real time. Stream data processing is the in-memory analysis of machine data that is in motion. The main purpose of the stream processing is to react to operational data analysis in real time. Spark, which is used in this work, is a general purpose computational framework providing more flexibility compared to MapReduce.

Turkish State Meteorological Service (TSMS) needed a weather-warning system based on the sensor data submitted by distributed climate sensor network around Turkey. In this work a prototype system was implemented to process the streaming sensor data using big data technology mainly Spark and produce weather alarm levels according to the meteorological specifications that would support TSMS's goal.

In this work, we propose a weather-warning system algorithm that uses Spark which is a fast and general processing engine compatible with Hadoop. Spark is designed to perform both batch processing and new workloads like streaming, interactive queries and machine learning. This big data technology provides highly scalable and fault-tolerant streaming processing while providing near real time processing.

The scope of the work was to demonstrate the tools and techniques to perform stream processing on rainfall and wind speed sensor data of TSMS and produce weather alarm levels accordingly. The exercise also aimed to integrate the dataset and the result set with the data discovery tool in order to present the findings of the analysis.

Current Data Architecture
There are 1700 Automatic Weather Stations (AWS) all around the Turkey and they observe and send 1 minute and 10 minute meteorological data continuously. In the near future, totally 3000 AWS are planning. The main communication method of the AWS is GPRS (General Packet Radio Service) but ADSL (Asymmetric Digital Subscriber Line) and VGAT (Very Small Aperture Terminal) technologies are also using. Only 140 station is manned and the others are unmanned. Data is collected by a TCP based socket software and archived in the Oracle database. At the end of the day, approximately 3 million data row inserted into database per day.

Archiving and retrieving databases are different. Archiving database is Oracle 12C for fast transactions, especially Insert. SAP Sybase IQ 16 is using for fast data retrieval process. SAP ETL tool is used for data transfer between the databases and while transferring data from Oracle to SAP Sybase IQ, some basic quality control processes are used.

The rainfall and wind speed data in our prototype was supplied by TSMS. The size of dataset was 4 GB and it included a total of 35 million records. The data was supplied by 18 different sensors located in Antalya-a southern region city in Turkey-for the period of 2010-2014.

Weather Warning System Business Rules
The weather warning system in Turkey uses color-coded symbols indicating different levels of severity and each level is associated with different criteria as shown below:

Yellow: This color means potential danger. The weather is unlikely to be extreme but care is called for in activities that are dependent on the weather.

Orange: This color means danger. There is severe weather that may cause damage or accidents. The weather brings risks. It is wise to be careful and keep abreast of the latest developments in the weather.

Red: Great danger from extremely severe weather. Major damage and accidents are likely, in many cases with threat to life and limb, over a wide area.

These criteria are taken from EUMETNET’s meteo-alarm system. In this work, two weather sensor data criteria for rainfall and wind speed were used. The color codes and the corresponding criteria are shown below:

The rain shower data was accumulated for every 12 hours and the accumulation level was compared with the thresholds given at the tables above. The system determined whether the rainfall level produced yellow, orange or red alarms. Similarly wind speed value was compared with the threshold values and the system produced the alarm levels correspondingly. Spark technology was used to process real time sensor data and produce alarm levels.

Identification Of Storm Detection Using Spark

Big Data Conceptual Structure

Execution Innovation

In this work, we propose a weather-warning system algorithm that uses Spark which is a fast and general processing engine compatible with Hadoop. Spark is designed to perform both batch processing and new workloads like streaming, interactive queries and machine learning. This big data technology provides highly scalable and fault-tolerant streaming processing while providing near real time processing.

Many companies use big data to make or facilitate user's decisions in the form of recommendation systems, predictive analysis etc. One of the key properties of any decision is latency that is the time it takes to make the decision from the moment the input data is available. Reducing latency can significantly increase the effectiveness of the analysis. Especially when these decisions are based on complex computational algorithms Spark is an ideal fit to speed up the critical decision making process. TSMS's sensor network data frequency and the corresponding requirement of processing that sensor data in real time fit into the streaming data processing scenario.

Results And Analysis
PoC process resulted in total of 8594 alarms for rain shower and wind sensor data. For rain shower data all three levels of alarm—yellow, orange, red- was produced whereas for wind data only yellow level was produced due to the data range. The following table displays the alarm level counts for the corresponding data.

Conclusions And Future Work
This experimental work successfully demonstrated the purpose of the exercise which was applying streaming data analysis on the meteorological sensor data and producing real time alarm levels according to the threshold levels determined by the department. Since the PoC has served its purpose on the two sensor data the streaming processing technology can be applied to the complete set of data to produce further real time analysis.

This work could be adapted to WIS data. There are many kind of meteorological data in WIS content and it might ensure more flexibility.

For the next steps of this project, it will need the large scale data from WIS web interfaces. WIS is operated by WMO community, built industry standards, incorporating existing services and solutions provided to public and WIS is a sharing platform which is providing some benefits. These are Accessibility, Interoperability, Visibility, Utility, Reliability etc.

Users access the all kind of meteorological data from WIS in different ways. Everybody find the data which they use for their project. This project may be develop by using the different kind of meteorological parameters. Because all WIS data can be managed and documented. They are discoverable, accessible and easy to use.