Smart Grids Data Management: A Case for Cassandra

Gil Pinheiro, Eugénia Vinagre, Isabel Praça, Zita Vale, Carlos Ramos

GECAD – Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development, Institute of Engineering, Polytechnic of Porto (ISEP/IPP), Portugal

{1090574, empvm, icp, zav, csr}@isep.ipp.pt

Abstract. The objective of this paper is to present a SMACK based platform for microgrids data storage and management. The platform is being used in a real microgrid, with an infrastructure that monitors and controls 3 buildings within the GECAD - ISEP/IPP campus, while, at the same time, receives and manages data sources coming from different types of buildings from associated partners, to whom intelligent services are being provided. Microgrid data comes in different formats, different rates and with an increasing volume, as the microgrid itself covers more customers and areas. Based on the atual available computational resources, a Big Data platform based on the SMACK stack was implemented and is presented. The Cassandra component of the stack has evolved. AC version 2 is still supported until the version 4 release, and is often still used in production environments. However, a new stable version, version 3, introduces major optimizations in the storage that bring disk space savings. The main focus of this work is on the Data Storage and the formalization of the data mapping in Cassandra version 3, which is contextualized by means of a short example with data coming from the monitoring infrastructure of the microgrid.

Keywords: Big Data Storage; Smart Grids; Cassandra

1 Introduction

Technological developments led to a huge spread of monitoring equipment that now provide an enormous quantity of data, based on which intelligent services may become available, turning into dynamic the traditionally centralized management of certain areas. In power systems, technological developments and the roll out of meters turn the Smart Grid (SG) as a new reality. Indeed, digital data sources range from sensors, that measure electric parameters (current, voltage, phase shift and frequency), to meters that monitor in real time consumption data and distributed generation sources, to environmental sensors (temperature, humidity, etc.), all of them being relevant to shift from a static structure to a more intelligent and flexible way to manage the electrical energy resources. The monitoring of the grid status results in a huge amount of collected data to deal and sharing with various parties [1].

The volume, velocity, and variety of the data make traditional data storage systems inappropriate to obtain the relevant value from the data analysis in a very short time.

S. Omatu et al. (eds.), *Distributed Computing and Artificial Intelligence, 14th International Conference*, Advances in Intelligent Systems and Computing 620, DOI 10.1007/978-3-319-62410-5_11

Big Data platforms are now the most promising way for the storing and analysis of high volumes of data. Apache technologies are being used in several domains. In this paper, we define a SMACK based architecture and implement a platform to support a real microgrid infrastructure existent at GECAD research group. Particular insights are given to the data storage process and the main focus of our contribution is given to the data mapping in relation to the partition size in Cassandra's most atual version.

The paper is structured into 4 sections, with section 2 addressing Big Data (BD) in SG context. Session 3 presents a platform based on SMACK, having Apache Cassandra (AC) as distributed storage for GECAD microgrid, with particular insights on the formalization of the data mapping process. Finally, the conclusions are presented in section 4.

2 Big Data and Smart Grids

To improve decision making, a system must be in place capable of collecting, managing, and processing information. In BD, the sheer volume of information requires new approaches when designing a solution that extracts knowledge within a reasonable period. This phenomenon, referred to as BD, is characterized by 5 Vs (i.e. Volume, Velocity, Variety, Veracity, Value) [2]. Each of these Vs represents real challenges (e.g. how to collect and transport a large volume of information; how to store this information, how to analyze and extract knowledge, how to ensure its security and privacy, how to process it in real time, etc.). The management of information with these characteristics raised great interest in the scientific and business community. Hadoop and Spark, are the most referenced frameworks.

The Apache Hadoop Framework was the first mainstream BD solution. It is based in Batch Processing, distributed file system HDFS (Hadoop Distributed File System), a programming model MapReduce and YARN (Yet Another Resource Negotiator) [3]. Apache Spark is a set of tools and high level APIs for large scale distributed processing of data in-memory [4]. Currently, Spark is considered as the most active open source project in BD. Its speed advantages, allied with an out of the box integration of data manipulation using SQL like syntax, support for several storage systems, and ability to distribute machine-learning computation, have contributed to its success.

In the new ecosystem of SG, all the players (i.e., power generation, transmission, distribution, customers, service providers, operations and markets) support their operations using a varied range of equipment that generate a large flow data. The last report issued by the European Union [5] refers numerous projects focusing the implementation of smart metering (SM). According to the same source, around 72 % EU customers are expected to be equipped with SM by 2020. The success of these projects launches an alert for the extensive amount of data generated in real time, that need adequate storage and analysis means to provide the development of dynamic services to better manage grid resources. Also, in the literature there are numerous references that characterize the type of data circulating on SG, at very high rates, as unstructured or, at most, semi-structured. Extrapolating this reality to the universe of the existent equipment and the foreseen roll outs by 2020 is easy to understand the