BIG DATA @ MICROSOFT

Raghu Ramakrishnan

Microsoft Cloud Information Services Laboratory
raghu@microsoft.com

Abstract. Until recently, data was gathered for well-defined objectives such as auditing, forensics, reporting and line-of-business operations; now, exploratory and predictive analysis is becoming ubiquitous, and the default increasingly is to capture and store any and all data, in anticipation of potential future strategic value. These differences in data heterogeneity, scale and usage are leading to a new generation of data management and analytic systems, where the emphasis is on supporting a wide range of very large datasets that are stored uniformly and analyzed seamlessly using whatever techniques are most appropriate, including traditional tools like SQL and BI and newer tools, e.g., for machine learning and stream analytics. These new systems are necessarily based on scale-out architectures for both storage and computation. Hadoop has become a key building block in the new generation of scale-out systems. On the storage side, HDFS has provided a cost-effective and scalable substrate for storing large heterogeneous datasets. However, as key customer and systems touch points are instrumented to log data, and Internet of Things applications become common, data in the enterprise is growing at a staggering pace, and the need to leverage different storage tiers (ranging from tape to main memory) is posing new challenges, leading to caching technologies, such as Spark. On the analytics side, the emergence of resource managers such as YARN has opened the door for analytics tools to bypass the Map-Reduce layer and directly exploit shared system resources while computing close to data copies. This trend is especially significant for iterative computations such as graph analytics and machine learning, for which Map-Reduce is widely recognized to be a poor fit. While Hadoop is widely recognized and used externally, Microsoft has long been at the forefront of Big Data analytics, with Cosmos and Scope supporting all internal customers. These internal services are a key part of our strategy going forward, and are enabling new state of the art externally facing services such as Azure Data Lake and more. I will examine these trends, and ground the talk by discussing the Microsoft Big Data stack.