As the Internet of Things and other data acquisition and generation technologies advance, data being generated is growing at an exponential rate at all scales in many online and scientific platforms. This mostly unstructured and variable data growing and moving between different applications dynamically in vast quantities is often referred to as "Big Data". The amount of potentially valuable information buried in Big Data is of interest to many data science applications ranging from natural sciences to marketing research. In order to analyze and digest such heterogeneous data, challenges for integration and distributed analysis include: scalable data preparation and analysis techniques; new and distributed programming paradigms; repeatable and verifiable process development; and innovative hardware and software systems that can serve applications based on their needs.

An important aspect of Big Data applications is the variability of technical needs and steps based on applications being developed. These applications typically involving data ingestion, preparation (e.g., extract, transform, and load), integration, analysis, visualization and dissemination are referred to as Data Science Workflows. A data science workflow development is the process of combining data and processes into a configurable, structured set of steps that implement automated computational solutions of an application with capabilities including provenance management, execution management and reporting tools, integration of distributed computation and data management technologies, ability to ingest local and remote scripts, and sensor management and data streaming interfaces. Each data science workflow has a set of technological challenges that can potentially employ a number of Big Data tools and middleware. Rapid programmability of applications on a use case basis requires workflow management tools that can interface to and facilitate integration of other tools. New programming techniques are needed for building effective and scalable solutions span across the data science workflows. Flexibility of workflow systems to combine tools and data together makes it an ideal choice for development of data science applications involving common Big Data programming patterns.

Big Data workflows have been an active research area since the introduction of scientific workflows. After the development and general adoption of MapReduce as a Big Data programming pattern, a number of workflow systems were built or extended to enable programmability of MapReduce applications including Oozie, Nova, Azkaban and Cascading. The Kepler Workflow Environment, developed by the WorDS Center of Excellence at SDSC led by Ilkay Altintas, also provide a distributed data-parallel (DDP) programming module on MapReduce and other BigData programing patterns on top of well-known Hadoop and Spark engines to build and execute big data workflows. The actor-oriented approach of Kepler provides flexibility and improves application programmability due to: (i) its heterogeneous nature in which Big Data programming patterns are placed as part of other workflow tasks; (ii) its visual programming approach that does not require scripting of Big Data patterns; (iii) its adaptability for execution of data parallel applications on different execution engines.

Marcar como concluído