According to Lin and Ryaboy analytics platform at Twitter has experienced tremendous growth over the past few years in terms of size, complexity, number of users, and variety of use cases. Twitter is a social networking site that stores lots of real-time data such as photos, videos, comments, etc. The evolution of our infrastructure and the development of capabilities for data mining on “big data”. It consists of more preparatory work that precedes the application of the data mining algorithm and is followed by substantial effort to turn preliminary models into robust solutions. There are two topics mentioned, they are: First, schemes play an important role in helping data scientists understanding the petabyte-scale data stores, but they are insufficient to provide the data available to generate insights. Second, a major challenge in building a data analytics platform stem from the heterogeneity of the various components that must be integrated into production workflows.

Introduction:

There are thousands of Hadoop nodes across multiple data centers. Each day, around one hundred terabytes of raw data are ingested into our main Hadoop data warehouse. Big data mining is about much more than what most academics consider data mining. A significant amount of tooling and infrastructure is required to operationalize vague strategic directives into concrete. Exploratory data analysis includes data cleaning and data munging not directly problems related to the hand. Data infrastructure engineers work to make sure that productionized workflows operate smoothly, efficiently, and robustly. This paper has two goals, they are:
1) For practitioners : To flatten bumps in the road.
2) For academic researchers: provide a broader context for data mining in a production environment. In addition, identify opportunities for future work that could contribute to streamlining big data mining infrastructure.

How we got here:

Three major trends distinguish insight-generation activities today, are: 1) First, a tremendous explosion in the sheer amount of data, orders of magnitude increase. 2) Second, we have seen increasing sophistication in the types of analyses that organizations perform on their vast data stores. Most of the information needs fall under online analytical processing (OLAP). Today’s data scientists are interested in predictive analytics i.e using machine learning techniques to train predictive models of user behavior. The application of data mining on behavioral data changes the scale at which algorithm needs to operate, and generally weaker signals present in such data requires more sophisticated algorithm to produce insights. 3) Third, open-source software is playing important role in the ecosystem. This includes many complimentary systems such as Hadoop, HBase, Hive, Pig, etc.

Big Data Mining Cycle:

Before beginning exploratory data analysis, the data scientist needs to know what data are available and how they are organized. They are some service architectures: 1) Service Architecture and Logging.
2) Exploratory Data Analysis.
3) Data Mining.
4) Production and Other Consideration.
5) Why should Academic care?

Schemas and Beyond:

An enterprise's most obviously valuable data, such as business objects representation customers, items in catalogs, purchases, contracts, etc. are represented in carefully designed schemas. The quality, usefulness, and provenance of these records are carefully monitored. Complex operations are often ad-hoc. A small team can be successful by using JSON. Some common points while building scalable big data infrastructure are: 1) Don't use MySQL as a Log. 2) The Log Transport Problem. 3) Plain Text to JSON. 4) Structure Representations. 5) Looking Beyond the Schema.

Plumbing:

One of the biggest challenges in building and operating a production analytics platform is handling impedance mismatches that arise from crossing boundaries between different systems and frameworks. Plumbing refers to the set of challenges.

Conclusion:

An efficient and successful big data analytics platform is about achieving the right balance between several competing factors: speed of development, ease of analytics, flexibility, scalability, robustness, etc.