Section 1: Week 1: Database Bibliography

Nate Bachmeier

TIM-8120: Distributed Systems

December 8th, 2019

North Central University

Bibliography

The growth of data creation from sources such as IoT, Cloud, Big Data, and Mobile (ICBM) is increasing at an exponential pace. This explosive volume of information is forming in different shapes, with varying degrees of structure. Traditional database patterns and practices are unable to manage these data sets efficiently, which is driving enterprise environments to invest in new technologies. Merely adding new widgets to the network topology will not solve the challenges, and existing business processes will also need a revision. Through a combination of these ideas, enterprises can evolve their data pipelines and unlock the insights into more agile data-driven decisions.

# Business Intelligence Tomorrow (2019)

To understand the landscape of future Business Intelligence systems, one needs to look at the challenges of today (Harper, 2019). Harper proposes investment areas into (1) metadata management, semantic understanding, data catalog, data modeling, (2) Natural Language Processing, and (3) Edge computing.

The first aspect deals with the Data Lifecycle Management of ICBM data and its operationalization. Storage prices have decreased significantly, which has led to numerous businesses collecting vast pools of unstructured dark data. One of the principal inhabitants for these businesses is a lack of Information Governance, which includes classification, controls, identification, and monitoring (Ajis & Baharin, 2019). As these organizational systems improve the usability and discoverability of specific data subsets, then data scientists can begin exploring the data and coming up with operational insights.

The second aspect deals with the interaction of users into the data management system. Many of these data repositories rely on query languages, such as Structured Query Language (SQL), to store and retrieve records. These languages introduce a barrier to entry challenges for users of the system, as they need to learn tedious syntax. Instead, Natural Language Processing (NLP) can convert business questions directly into data-driven solutions. When the NLP algorithm is made aware of proprietary object models, it can derive entities, verbs, and other relationships. These capabilities led to a democratization of self-service scenarios across all levels of the organization.

The third aspect is the inclusion of edge computing in data processing architectures. Micro-clouds of IoT and mobile devices are generating vast collections of sensor and machine-to-machine data. Centrally processing these feeds could involve significant network I/O, or is economically prohibitive to move. Instead, a transformation of these ‘high volume/low quality’ feeds into ‘low volume/high quality’ aggregations needs to take place. For instance, one hundred temperature sensors could report one hundred individual measurements or the median of their aggregate value.

# Big Data Quality: A Survey (2018)

Big data is high-volume, high-velocity, and high-variety information that needs to be transformed into high-quality models (Taleb et al., 2018). According to ISO 25012, the critical measurement of quality is its fitness for us. Taleb et al. measure this fitness against intrinsic, contextual, accessibility, and representational dimensions. They propose a Process-Driven Analytic Pipeline that contains multiple quality gates. Each quality gate further refines the information into a more standardized and curated representation of itself. For instance, unstructured data becomes annotated, reformatted, and entities extracted. This curation process improves the consumption of data into analytical solutions, enabling more precise estimations.

Their approach is similar to the National Institute of Standards and Technology (NIST) reference model for data lifecycle management, which includes phases Collection, Preparation, Analysis, and Action (Mazumdar, Seybold, Kritikos, & Verginadis, 2019). NIST also proposes spending effort upfront to clean data sets. For instance, extreme outliers might skew the prediction results, and by pre-emptively removing them, a more accurate model defined.

# Business Intelligence for Enterprise Systems: A Survey (2017)

After data curation, experts use machine learning to perform statistical analysis (Duan & Xy, 2017). Many supervised algorithms, such as linear regression, exist to associate data features with a known label. An example dataset might contain housing sales prices, along with its age and square footage information. Given