Paul M. Washburn
CSCI E-11
Unit 1: Big Data
October 6, 2019

Future Technology: Office Networking Application

There are numerous technologies covered in this unit on Big Data that are likely to be crucial in implementing a successful Office Networking App. Each will be discussed in-turn below.

One of the most exciting technologies to be employed in the implementation of an office networking app will be effectively leveraging modern tools to wrangle & unite all of the disparate data sources relevant to the modern office environment. Important examples of these data may include popular communication platforms such as Slack or Microsoft Office – for which integration may be automated via API. However, dealing with the "long tail" of alternative sources (such as information transcribed from audible conversations, physical fascmile, or company cell phone text messages) may be tougher to integrate. The same idea is present in mining the web for "business reputation" indicators. All this data will need to be integrated in order to be maximally useful to the user. Data can be integrated similarly to some of the applications discussed in class – via some data concatenation, normalization, and distance from other points (via cosine similarity or some other measure of distance). Getting the data into one centralized source will be essential to the success of this application.

Cloud technologies of all breeds will be a cornerstone in the success of the proposed venture, especially as they relate to storage/architecture and analytics. A columnar storage model will be preferred for all database storage to enable efficient retrieval and storage space. An ArrayDB-like system will be employed for efficient ML-OPS, such that a query and a complex analytical operation (such as Singular Value Decomposition) can be performed in one job. This will be vital for supporting tens of thousands of models. Performing search that is "embarrassingly parallel" (such as searching for a specific topic in a user's conversations using an NLP-based search mechanism) will be the only area where something like Hadoop will be employed such that the work is distributed over many workers. In the beginning, much of the exploratory analysis and large-scale training of ML models will be performed using Spark to take advantage of main memory computations. The same memory-first approach will be taken with the firm's OLTP database for housing the firm's transactions.

Some of the useful data for this application will be of high velocity, and in order to make timely use of it there will need to be some sort of approximation in computation & simplification in storage. For example, analyzing graph data of users' calendar appointments with colleagues will contain a lot of noise (e.g. any given meeting typically won't amount to much), but identifying higher order patterns from unrelated users may help shed light on the

health of inter-colleague relationships. This invites the use of coresets for representing data. This will allow for minimal error in computations at the benefit of timely answering of important analytical questions. Some data of this sort will be dealt with via streaming (summarized yet fully evaluated) while some will be dealt with via sampling when outliers are not prevalent.

Investigating data for intuitive understanding is essential for actionable business understanding. That is why interactive data visualization will be allocated more resources for this office networking app than is typical in similar organizations. Allowing users to perform direct manipulation on the data via a user interface that centers upon a visualization of said-data helps facilitate such understanding. Users will be able to pan, zoom, filter and facet the data in real-time while they investigate hypotheses. Of course, visualizations will be designed to have an appropriate graphical integrity ration (near 1) and optimized for users.

Machine learning of all sorts will be central to this application. Plain vanilla classification systems will be used where appropriate (and supported where appropriate with big data tools), yet so will regression and NLP systems. Multi-aspect summarization is another interesting technology that can be applied to the proposed Office Networking App and will be explored in an R&D context for this task. Unsupervised labeling of sentences (whether from Slack or from an in-person meeting) might help a user identify meaningful conversations & changing relationships, such as when a conflict may have arisen. Implementing recommendation systems will also be enabled by implementing co-occurrence matrices for users vs. features, and inform recommendations for actions to improve their relationships & outcomes at work. Property testing will be helpful in the acquisition of a representative sample for use in training machine learning models when there is so much potential data that the noise-to-signal ratio is an impediment.

As the data generated per person continues to grow at an exponential pace, the big data technologies covered in this module will be of ever-increasing importance to the successful implementation of the proposed office networking app. It is likely that more research into related technologies will bear fruit as well.