**Requirements analysis and motivation**

The main motivation and the actual requirements specifications of the proposed researched innovation is not derived from potential customer needs nor from potential use-case scenarios but rather from actual existing customers that are using our innovation in a production-grade system. Our proposes innovation that is presented within this thesis is part of a large scale predictive analytics system that has multiple separate pipelines ranging from our real-life use case - that of real time preparation of market basket analytics based on tensor graph parallel computation, customer retention predictive analytics models (done by other junior data scientist student), recommendation engines, event prediction. All these modules have been researched and developed together with an additional ETL team of two engineering and the overall supervision of a senior data scientist (PhD candidate). It is mandatory to mention that our real-life use-case is the main source for our proposed research and innovation presented in this thesis. Even more, the presented research resulting innovation is totally mandatory for the actually functioning and commercial application of our production grade use-case.

The simplest approach to explaining the proposed innovation motivation is the presentation of the actual process that generated the need for the innovation. The actual real-life production grade use-case of our systems involves, as previously mentioned, the parallel processing of massive amounts of data using tensor computational graphs. Basically, our real-life goal is to provide in a short amount of time the final user/customer of the system with various analysis of the available commercial transactional information, thus constructing a complete relational graph between any two different entities. In order to have even more in-depth information, we can imagine a pharmaceutical retail company that needs a product similarity analysis, product complementarity matching, products composition in encapsulated products, decomposition of compel products, affection-based micro or meta clustering of all its products - all of these features based entirely on customer transactional behavior without any other features. For this hypothesis we have researched and developed our own Deep Learning models (as presented in our research paper „..........”) that use NLP-related method in order to generate, with two different approaches, six different products „meta-maps”. As stated, detailed information about our research in the area of Deep Learning for Predictive Business Analytics is beyond the scope of this thesis and can be further analyzed in the mentioned paper, however we will briefly explain the real-life end-customer problem that resulted in the need for our proposed innovation in a step-by-step approach:

1. The end-user of the commercial system is presented with the feature of our system: that of constructing advanced analytics for the so called classic „Market Basket Problem”
2. In order to produce a proof-of-concept in real-time and provide the customer with the confidence that our product delivers promised feature we make the analysis on-the-spot without any cloud-backed resources, without any process of ETL and without any need for customer data export to our internal infrastructure.
3. For the previous mentioned simple proof-of-concept we use fully confidentiallized transactional data that contains only TIMESTAMP, TRANSACTION\_UNIQUE\_ID and PRODUCT\_ID
4. By employing our own models based on tensor computational graph models we have to create several latent vector multi-dimensional spaces where the products are defined with 128 numbers based on 32 bit floats. The generated latent multi-dimensional spaces can then be used to produce the previously mentioned end-user analytics.
5. In order to run the previously defined experiment we need an environment that will be capable of executing multiple graph computation with multiple data-sources in parallel in order to minimize the experiment running time. Due to the nature of the required tensor computations, basically large matrix multiplications and additions, it is obviously required to employ a simple setup of massive parallel computing that could be used based on mobile computing infrastructure. Thus, we use a basic setup of minimum 386 numerical computing cores provided by a GPU device with a 2 GB RAM. At this point we can define two hypotheses:
   1. We must minimize the execution of our tensor computation by using the array of numerical cores
   2. We must minimize the total execution time of our whole experiment by “squeezing” all the tensor computational graphs and required input data within the available infrastructure

Finally, at this point we arrived at the conclusion that a special GPU parallel computation allocation engine is required that will augment the existing capabilities of the state-of-the-art tensor computation frameworks (TensorFlow in our particular case) with the proposed memory and computation “squeeze” feature. This basic requirement resulted in the following main analysis objectives that provided the fundament of our proposed innovation:

[