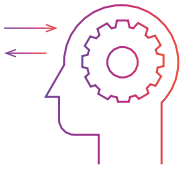
Enabling Next Generation Enterprise IT - EIT 2.0

Accelerating development and deployment of applications in EIT 2.0

WHITE PAPER



Abstract

With the rapid progress in machine learning (ML) research, we have noticed a paradigm shift in application development from traditional control-flow Software 1.0 to data-flow Software 2.0 programming. This usage of ML-based models has replaced customer scoring techniques for generating recommendations. Software 2.0 programming is data-driven and requires a vast amount of clean, labeled, unbiased data sets, and a clearly defined architecture for neural network models, efficient techniques for model training, rigorous testing, and high-performance deployment.

Since Software 2.0 programming invites heterogeneity in the entire application lifecycle – from development to deployment – various architecture and performance-related challenges arise. This paper outlines the requirements for migrating parts of Software 1.0 to the Software 2.0 framework and the challenges associated with accelerating the development and deployment of Software 2.0. In envisioning the evolution of existing enterprise IT systems to an agile, collaborative, and data-driven enterprise IT system (EIT 2.0), we compare the conventional application lifecycle development approach with that in EIT 2.0. We also highlight the challenges one needs to address where both Software1.0 and Software 2.0 coexist in an enterprise IT system.

TATA CONSULTANCY SERVICES

AI for All

A scalable deployment stack of enterprise applications consists of multiple layers of cluster management, data management, network management and application servers, referred to as an enterprise IT system. Traditionally, the software in these components is built using task-driven programming paradigms, where business requirements are mapped to an executable task using a set of functions. However, with an exponential increase in ML and DL research, such complex functions are being replaced with data-driven models that help enterprises better while reducing human intervention.

A modern enterprise IT system (EIT2.0) employs these data-driven models to automate some of the tasks as below:

- A data management system may use reinforcement learning or recurrent neural networks to decide an optimal plan for query execution.
- Data structures such as B-tree indexes may be replaced by models trained on the data access pattern.
- Resource provisioning in a cluster may be automated by observing workload patterns and their respective resource consumptions in a model.
- Core functionalities of network management such as traffic prediction, routing and classification, congestion control, resource and fault management, QoS and QoE management, and network security can be automated using ML/DL algorithms.

Tim Kraska and his fellow MIT scholars have envisioned a new type of data processing system, SageDB, a radical approach to building database systems by using ML models combined with program synthesis to generate system mponents that are driven by an AI engine.

Democratizing AI to Increase Engagement and Realize Benefits

One of the main benefits of using data-driven trained models in businesses is making intelligence available to everyone across the enterprise, including non-programmers like data scientists. Democratizing AI by building frameworks and/or systems that are accessible to all lowers the entry barrier and increases engagement. This accelerates data-driven application development by abstracting the underlying system complexity and hardware heterogeneity through the support of high-performance libraries to be used by data scientists. Also, some tools facilitate building models using domain functions, like Snorkel from Stanford, that can be used to label the large-sized training data required for DL algorithms. The availability of a democratized AI framework encourages users to replace task-driven functions with data-driven models. For example, a solution architecture for a recommendation system may use statistics to identify co-relations across different data sets, analyze user behavior, and build a mathematical scoring method to recommend a product to a user. However, democratized AI frameworks can automatically build the recommendation model with the required data sets.

Modern applications, composed of both data-driven models and task-driven functions, complicate the process of testing. The data-driven models are sensitive to any change in input data, whereas the accuracy of task-driven functions depends on the control-flow coded in a function. The test cases must be designed to capture these sensitivities. We, at TCS Research, have designed a process, similar to the waterfall model in traditional software development, for developing such applications with appropriate feedback loops[1].

1. *Rekha Singhal et al. 2019. Fast Online 'Next Best Offers' using Deep Learning. In Proceedings of the ACM India Joint International Conference on Data Science, and Management of Data, COMAD/CODS 2019, Kolkata, India, January 3-5, 2019, 217–223.*

The TCS Way

Data management is the fundamental problem in data-driven programming. Supervision learning works with labeled data, which is a rare commodity. Snorkel[2] labels the data using the domain rules. We have extended this to auto-label data using small amounts of labeled data and also recommend the right amount of labels needed for the desired accuracy.

We have further observed about the development of many ML pipelines on the same data sets for different use cases for one client. Can these pipelines synergize with each other by reusing features for models computed across different pipelines? Yes, it is possible; by designing a high-performance feature store, which can be reused by data scientists across multiple ML/DL pipelines. For example, a client may require building a personalized home page as well as next best offer for a user. In this case, the same data sets may be used for both pipelines.

This high-performance feature store is scalable and configurable to a client's requirements and is optimized for memory size and access latency.
The use of AI to accelerate data access by replacing complex data structures (such as K-d tree) and algorithms (such as sorting) with data-driven models can also be explored.

- Deployment of data-driven models employs heterogeneity in the whole development stack and facilitates:
- Integration of heterogeneous data sources (e.g., text, relational, images),
- Interoperability of streaming engines (e.g., flink, ignite),
- Interoperability across model building languages (e.g., R, Python),
- Interoperability across models (e.g., tensorflow, pytorch),

2. *Alex Ratner, Braden Hancock, Jared Dunnmon, Roger Goldman, and Christopher Ré. 2018. Snorkel metal: Weak supervision for multi-task learning. In Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*

- Integration of data processing engines (e.g., Oracle, Timesten, Textual store),
- Integration of different file systems (e.g., S3, HDFS),
- Embracing different compute (e.g., CPU, GPU, TPU, FPGA) and memory (persistent memory, NVRAM, Flash) for high performance.

There have been many solutions[3,4,5,6,7] in the past that address this heterogeneity at different levels, but the ask is to leverage this in the whole stack for optimal performance. This approach has been studied in depth in collaboration with Stanford. It envisions a new kind of system, Polystore++[3] for embracing and accelerating access to heterogeneous data processing engines using heterogeneous hardware accelerators. There is a need to have benchmarks for such applications to decide the right technology stack for deployment. For this also, models[8] have been developed to predict task-driven application performance for big data systems. Further, performance prediction models are being mapped data-driven models for estimating their training and inference time on heterogeneous hardware including clouds services.

---

3. *Rekha Singhal, Nathan Zhang, Luigi Nardi, Muhammad Shahbaz, and KunleOlukotun. 2019. Polystore++: Accelerated Polystore System for Heterogeneous Workloads. In 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). IEEE, 1641–1651.*
4. *Apache Beam 2016. Apache Beam: An advanced unified programming model. Retrieved Jan 01, 2020 from https://beam.apache.org/*
5. *Arrow 2016. Apache Arrow: A cross-language development platform for in-memory data. Retrieved Jan 01, 2020 from https://arrow.apache.org/*
6. *W Lin et al. 2019. ONNC: A Compilation Framework Connecting ONNX to Proprietary Deep Learning Accelerators. In 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS). 214-218. https://doi.org/ 10.1109/AICAS.2019.877151*
7. *David Koeplinger et al. 2018. Spatial: A Language and Compiler for Application Accelerators. In ACM PLDI.*
8. *Manoj K. Nambiar, Ajay Kattepur, Gopal haskaran, Rekha Singhal, and Subhasri Duttagupta. 2016. Model Driven Software Performance Engineering: Current Challenges and Way Ahead. SIGMETRICS*

Conclusion

The emergence of a new ecosystem of ML/DL-driven applications, or Software 2.0, powers our vision of an agile EIT2.0. However, it is imperative to understand and accept instances where Software 1.0 is also present and therefore the ecosystem must adjust and address challenges where both models co-exist. Democratizing AI will help mitigate performance challenges and drive innovation at scale.

About The Author

Rekha Singhal

Rekha Singhal is a Principal scientist and heads the Computing Systems Research area in TCS. Her focus is on the areas of accelerating the development and deployment of enterprise applications in data-driven programming environment. Her research interests include heterogeneous architectures for accelerating ML pipelines, Learned systems, high-performance data analytics systems, big data performance analysis, query optimization, storage area networks, and distributed systems.

A senior ACM member, Rekha has several patents and publications in international and national conferences, workshops, and journals.

Rekha received her M.Tech. and Ph.D. in Computer Science from IIT, Delhi, and has been a visiting researcher for a year at Stanford University, United States.

Contact

Visit: TCS Incubation, https://www.tcs.com/tcs-incubation

Blogs: Research and Innovation, https://www.tcs.com/blogs/research-and-innovation

Email: tcs.incubation@tcs.com

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that has been partnering with many of the world's largest businesses in their transformation journeys for over 50 years. TCS offers a consulting-led, cognitive powered, integrated portfolio of business, technology and engineering services and solutions. This is delivered through its unique Location Independent Agile™ delivery model, recognized as a benchmark of excellence in software development.

A part of the Tata group, India's largest multinational business group, TCS has over 453,000 of the world's best-trained consultants in 46 countries. The company generated consolidated revenues of US $22 billion in the fiscal year ended March 31, 2020, and is listed on the BSE (formerly Bombay Stock Exchange) and the NSE (National Stock Exchange) in India. TCS' proactive stance on climate change and award-winning work with communities across the world have earned it a place in leading sustainability indices such as the Dow Jones Sustainability Index (DJSI), MSCI Global Sustainability Index and the FTSE4Good Emerging Index.

For more information, visit us at **www.tcs.com**

Experience certainty.

IT Services
Business Solutions
Consulting

Corporate Marketing | Design Services | M | 04 | 21