

Sustainability TechSprint 2021: TechSprint Demos

Thursday, 21 October 2021

ESGATA

Presenter: Sharon Lee, Catalin Rus

Judges: Francesca Hopwood Road (Chair)

- Bruce Harrison; Managing Director, BlackRock
- Elena Strbac; Global Head, Data Innovation, Standard Chartered
- Emily Keaney; Director of Legislative Reform, Information Commissioner's Office
- Emily Pierce; Assistant Director, International Affairs, U.S. Securities and Exchange Commission
- Jayne Wickins; UK Banking Director, Microsoft
- Stephen Ingledew; FinTech Scotland Executive Chair
- Satoshi Ikeda; Chief Sustainable Finance Officer, Japan Financial Services Agency

Catalin Rus

Good afternoon everyone on behalf of team 6. Thank you for attending. Next slide please. Team 6's specializations are focused on our deep understanding of systems and data, machine learning, solutions architecture and software engineering. Sharon Lee and myself will present our recommendations for use case 2. As the UK is preparing to host the COP26 UN Climate Change Summit, and following announcements of increased government funding for environmental business, there is a strong need to capture the wealth of unstructured data from ESG disclosures and other sources to help determine the need for any further regulatory intervention. Next slide please

We see the issues presented as a combination of processing and architecture, therefore, we'd like to propose a solution which includes data processing mechanisms and the architecture to support them. I'll discuss the high-level architecture before handing over to Sharon to discuss the machine learning aspects. We conclude that the system is dealing with structured and unstructured data of unknown volume with uncertain data quality and applicability. Given the unpredictable volume of data, we'd recommend a scalable, flexible solution that makes use of proven machine learning technology in the cloud. AI and machine learning would be used to process the unstructured data. Next slide please. Development would use industry standard design patterns, agile development infrastructure as code, continuous integration, unit testing and UAT. No single aspect of the architecture is immutable, and each may be substituted for another component. The system would consist of three basic modules: input progressing and output, each would be loosely coupled and is modularized for flexibility, using loosely coupled components, for example the cloud platform can be switched to on-prem microservices or another cloud provider. Loosely coupled processes will be used, lambda functions on demand, compute instances with other components, auto-scaling on demand. Input stage - in order to onboard data sources, either structured or unstructured, we propose a 5GL low code/no code input stage, which reduces the degree of the involvement from the professional developers and enables data analysts to evaluate data more easily. Optimized software may involve development effort on a cost benefit basis however, the initial data feeds set up and reviewed work will be mostly

completed at the input stage. End points will be auto-scaled so whether the system can consume IoT sensor data or a folder with thousands of reports, the data can flow securely to the processing engine with low latency and high availability. Next slide please.

For the processes, machine learning and cloud computing resources, discussed in more detail later, will be leveraged to analyse data before the data flows, next slide please. The output stage again, decoupled, we would make general proposals for data storage systems based on anticipated usage. At a glance data and dashboard visualizations would provide a front screen which leads into a Tableau front end, such as the angular grid, enabling large records, set traversal and a high-quality user experience. Therefore, we have a system that is flexible, scalable, and responsive to change with a low-cost barrier to entry. That concludes the architectural side of the presentation. And I'd like to hand over to Sharon to discuss machine learning.

Sharon Lee

Thank you, next slide please. so one of the main challenges for the system is how we decouple ESG data from different sources including the restructured data from commercial data providers and the semi-structured data and unstructured data, for example the company internal data documents. Any reports are publicly available, with news articles and even including social media, so we propose to use machine learning, and particularly natural language processing so that it can help us to quickly identify the key words and relevant information, and the values can be put into some, predefined structured data or table. We also want to calculate the confidence score based on the availability of the data sources. The processed data and the confidence score will then put into some predefined format, the standards and the matrix will come from the regulators very soon, hopefully. Next slide please. So, all data will be stored in the cloud infrastructure after the data acquisitions, and a detailed level of data quality check will be conducted to detect some contradictory data from different providers. There will be options to automatically detect the problematic data from the data providers. Administrators will also have the options to include or exclude the low confidence data. At the output we also want to support the development of self-created modules, dashboard visualisation as well as data analytics, including statistics, benchmarking or some additional machine learning capabilities such as anomaly detection, which can help the regulator to find the problematic data and carry out further investigation. Next slide please. That's all basically, thank you very much.

Francesca Hopwood-Road

Thank you very much indeed, Team 6. We will go straight to a question that's hot off the press from Jayne.

Jayne Wickins

Thank you team, how you will present the data and findings?

Sharon Lee

So we will have some standard dashboards, some predefined visualisation, which hopefully will be helpful to the regulators and some investors. However we will try to support some customizations such as Python coding and additional data analytics. We will also allow the user to do some SQL queries over the centralized data set.

Francesca Hopwood-Road

Thank you and perhaps I could ask a question. So given that unstructured data like social media data is increasingly visual and video based, would the tool be able to capture and analyse images and like video as well?

Sharon Lee

To start with I would say texts related information is easier to process. However based on the advancements in deep learning and also cloud computing in the near future, I'm sure we will have the computational power to process videos and images we would like to. Before we do that we would like to know about how much content and the entry of the information, because its quite computationally expensive but yeah, technically it is able to do that.

Francesca Hopwood-Road

Thank you. Elena your question please?

Elena Strbac

Thank you, so you mentioned about end users running SQL queries, so just to expand on that, how technical do you think your end users would need to be to use the system?

Sharon Lee

I would say it will require a little bit of background knowledge but we can implement a user interface so that the user can click in some selected columns and try and do some basic statistics and calculations I would say the basic functionality will be there but if the user wants to do advanced operations over the data then they will need to have some technical background to actually do it.