นางสาวอรปรียา ศรีทราย รหัสนิสิต 6230301067

สรุป Data / AI Summit 2021

ในปี 2013 Data Science ส่วนใหญ่ทำงานบน Scala ร้อยละ 92 แต่ในปี 2021 Scala ถูกใช้งานลดน้อยลง และใช้ Python และ SQL เพิ่มมากขึ้น ทำให้มีการจำลองเป็นเครื่องเซิร์ฟเวอร์ของเครื่องคอมพิวเตอร์เพิ่มขึ้น และใช้ library ต่างๆมาจัดการข้อมูล ที่มีขนาดข้อมูลไม่เกิน 1 GB แต่ในความเป็นจริงข้อมูลมีมากกว่า 1 GB จึงมีการพัฒนาเป็น 1 TB และ library ที่สำคัญของ Data Science คือ pandas ซึ่ง Spark สามารถทำงานได้มากกว่าใช้งานบนเครื่องคอมพิวเตอร์เพียงเครื่องเดียว จึงมีการพัฒนา pandas เป็น Koalas เพื่อให้สามารถใช้บน Spark ได้ หลักการทำงานเดียวกับ stack overflow ในการทำงานของ Koalas เป็นการดำเนินการบน pipeline ข้อดีคือสามารถปรับขนาดและประสิทธิภาพของ pyspark ที่ทำงานได้ตั้งแต่ 1 โนดและข้อมูลจำนวนมาก การทำงานจึงดีกว่า pandas และ Koalas สามารถทำ Visualization เหมือน pandas ได้ แต่ในการอ่าน error ของ pyspark ยังเป็นเรื่องยาก จึงมีโปรเจ็ค zen ที่พยายามลด error จากหลายบรรทัดให้น้อยลงและอ่านเข้าใจได้ง่ายขึ้น จากการศึกษาวิจัยจากบริษัทรถยนตร์ และจากผู้ใช้ที่จัดการข้อมูลที่พักเพื่อวิเคราะห์การวางแผนวันหยุดบนเกาะฮาวาย เปรียบเทียบการทำงานของ pandas บน pyspark ที่มีการรวม koalas ไว้ ที่ง่ายต่อการใช้งานกับ python และ SQL

MLflow เป็นแพลตฟอร์มที่เรียนรู้เกี่ยวกับ machine learning จากในปี 2011 software มีบทบาทในบริษัทรวมทั้งอุตสาหกรรมต่างๆเพิ่มขึ้น และเริ่มมี Ai ที่เข้ามาใช้งานในซอฟต์แวร์มากขึ้นเช่นกัน เมื่อไม่นานมานี้ GPT-3 ได้นำข้อความต่างๆมาใช้สร้างเป็นรูปภาพ และ google Ai มีงานวินิจฉัยทางการแพทย์ เช่น การวัดระดับฮีโมโกลบินในผู้ป่วยโดยการสแกนผ่านเรตินา รวมถึงการใช้ฝึกหุ่นยนตร์ให้มีความสามารถในการทำงานด้านต่างๆ บริษัทที่ใช้ Ai ในปัจจุบันคือ Facebook (fb learner) , Google (TensorFlow), NETFLIX และ Uber (Michelangelo) ซึ่งหลักการพื้นฐานของ MLflow มี 4 ส่วนองค์ประกอบคือ tracking, project, models และ model registry หรือก็คือ เทรนแล้วติดตั้งข้อมูลหรือ Deployment , นำข้อมูลดิบที่ได้มาใช้ในโมเดลและทำการตรวจสอบข้อมูลหรือ Data prep โดยมี model registry เป็นตัวกลางเพื่อนำไปเทรนต่อไป และใน MLflow มี mlflow.kares.autolog() เป็นเครื่องมือที่บันทึกข้อมูลที่เกี่ยวข้องกับโมเดลทั้งหมดอัตโนมัติ

Pytorch คือการทำงานด้วย PyCaret ที่เป็น library เกี่ยวกับ machine learning นำมาทำงานร่วมกับ MLflow สำหรับในการตั้งค่าความสามารถต่างๆ และให้ MLflow บันทึกค่าข้อมูลให้อัตโนมัติ

Roadmap คือ Captum library ใช้ร่วมกับ Pytorch ช่วยในเรื่องการวิเคราะห์ตีความด้วยแบบจำลองที่หลากหลาย

ในปี 2019 Ai เข้ามามีบทบาทในการทำงานของซอฟต์แวร์มากขึ้นพร้อมกับการใช้ข้อมูลมาทำงานร่วมด้วยเพื่อวิเคราะห์ข้อมูลให้แม่นยำมากขึ้น แต่ Ai แต่ละแบบจะมีประสิทธิภาพที่มากหรือน้อยขึ้นอยู่กับข้อมูลนั้นๆ อีกทั้งการนำ Ai มาใช้จะยากกว่าการโค้ดทางด้านซอฟต์แวร์ เพราะต้องนำทั้งข้อมูลและการโค้ดมาทำงานร่วมกัน

Pycarat/AutoML เป็น low-code ML ช่วยจัดการเรียกใช้ ทดสอบและเปรียบเทียบผลลัพธ์ของMachine Learning Model แต่ละโมเดล ประกอบด้วย Data , Target, Classification/clustering , computation

Feature Stores อยู่ใน Databricks ML platform เก็บและแชร์ Feature และสามารถนำ Feature ของคนอื่นมาใช้กับโปรเจคเราได้ แต่มีค่าใช้จ่าย

การทำงานของ Ai จะทำงานเป็นเลเยอร์ คือ นำข้อมูลมาเรียนรู้ที่ละชั้น เช่น IMAGENET ที่นำภาพหลายๆภาพมาเรียนรู้และต้องการข้อมูลปริมาณมากในการเรียนรู้ ในกระบวนการของ Ai จะใช้ ข้อมูลหรือ data , เครื่องคอมพิวเตอร์จำนวนมหาศาลในการคิดคำนวณ และบทบาทของ Ai จะมีเพิ่มมากขึ้น ซึ่งช่วยเปลี่ยนแปลงโลกในด้านพลังงานสะอาด และความเสมอภาคได้

ศึกษาเพิ่มเติม Data-native, Collaborative, Full ML Lifecycle Solution