TechNetCyber
May 2-4, 2023 · Baltimore, MD

Booth #2419

HITACHI
Inspire the Next

The Urgency of Action: Focused, Aligned and Ready

SIGNAL Innovation Showcase

Data Enrichment using LLM Text Analytics

Pragyansmita Nayak, Ph.D
Chief Data Scientist, Hitachi Vantara Federal
May 2023

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Speaker Intro

- Chief Data Scientist at Hitachi Vantara Federal (HVF)
  - Explores the "Art to the Science" of solution architectures orchestrating data, APIs, algorithms and applications enabling operational data fabric and data management solutions
  - Focus on Federal Government domain data for ~20 years
  - AFCEA Member since 2018
- Ph.D. in Computational Sciences and Informatics, GMU
  - Machine Learning Galaxy Redshifts using Color and Morphology
  - Dataset : Sloan Digital Sky Survey (SDSS)
- Enjoys cooking, travel and photography

LinkedIn : https://www.linkedin.com/in/pragyansmita

Twitter : https://twitter.com/SorishaPragyan

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Abstract

Text Analytics is a branch of machine learning (ML) combined with natural language processing (NLP) which extracts key characteristics from unstructured text documents. The additional processing cycles can improve trust for data assets by expanding their accessibility. Improved search of the content through correlation statistics, sentiment analysis, entity recognition and summarization can be easily automated.

Using Large Language Models (LLMs) such as OpenAI's ChatGPT, this talk will explore the potential data engineering operationalization and its impact on the associated metadata management.

Keywords LLM, ChatGPT, Text Analytics, Data Engineering, Unstructured Data, Sentiment Analysis, Summarization, Metadata Management, Feature Engineering, Hitachi Lumada DataOps

https://events.afcea.org/AFCEACyber23/Public/SessionDetails.aspx?FromPage=Sessions.aspx&SessionID=10230&SessionDateID=702

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Hitachi Vantara Federal’s Strengths

Protection

Safeguard and protect your most valuable asset – data.

Recovery

Detect anomalies and recover more quickly to keep mission continuity a priority.

Insight

Build intelligence across your entire data ecosystem to mitigate risks, control governance, and improve data quality and trust.

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

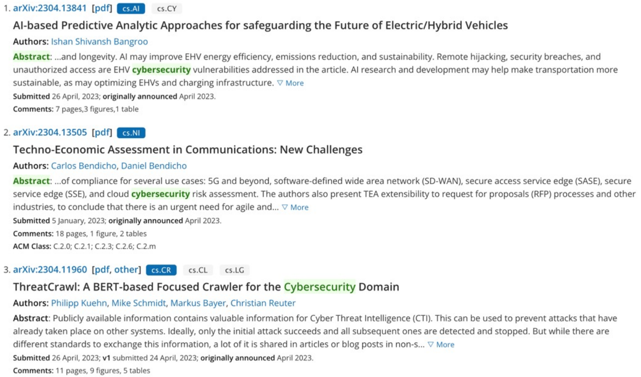
Text Analytics

- Decision intelligence multiplier from content, structure and patterns of textual unstructured/semi-structured data and metadata
- Leverages natural language processing (NLP), machine learning (ML), and computational linguistics to process and analyze text data at scale

Hitachi Vantara Federal @HVFederal · Apr 28

Join us next week at #TechNetCyber in Baltimore, MD! Stop by Booth 2419 and get a glimpse into our framework for #cyberresiliency and #dataprotection. Learn more: hitachivantarafederal.com/about/events/t …

- ✓ Automated extraction of information
- ✓ Identification of key concepts
- ✓ Sentiment analysis
- ✓ Topic modeling
- ✓ Entity recognition
- ✓ Other linguistic and statistical analysis

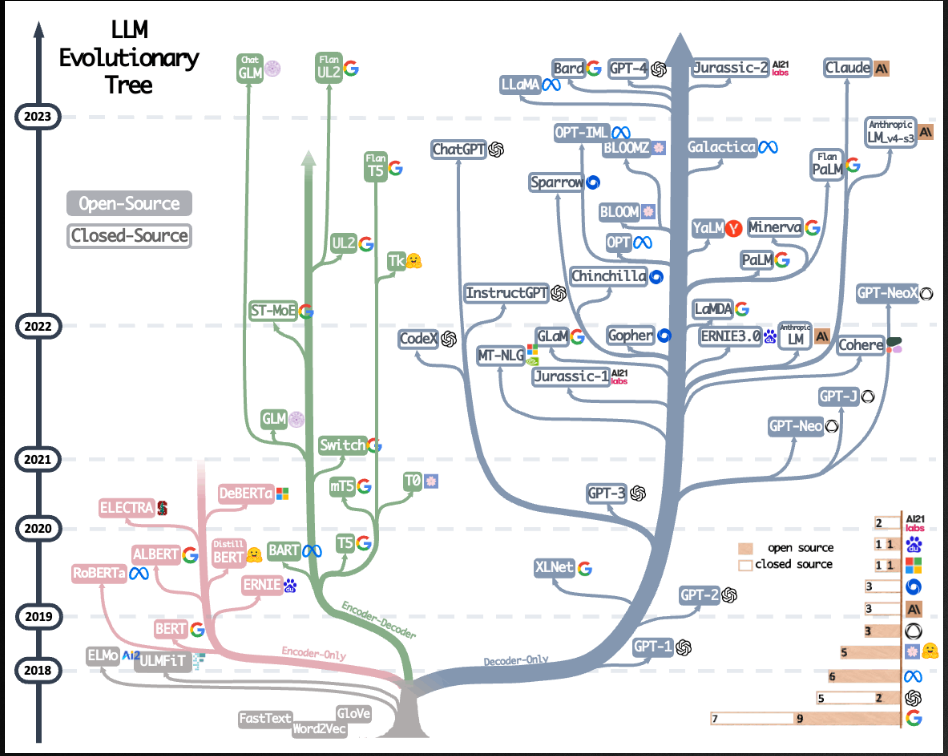


1. arXiv:2304.13841 [pdf] cs.AI cs.CY
   AI-based Predictive Analytic Approaches for safeguarding the Future of Electric/Hybrid Vehicles
   Authors: Ishan Shivansh Bangroo
   Abstract: ...and longevity. AI may improve EHV energy efficiency, emissions reduction, and sustainability. Remote hijacking, security breaches, and unauthorized access are EHV **cybersecurity** vulnerabilities addressed in the article. AI research and development may help make transportation more sustainable, as may optimizing EHVs and charging infrastructure. ▽ More
   Submitted 26 April, 2023; **originally announced** April 2023.
   Comments: 7 pages,3 figures,1 table
2. arXiv:2304.13505 [pdf] cs.NI
   Techno-Economic Assessment in Communications: New Challenges
   Authors: Carlos Bendicho, Daniel Bendicho
   Abstract: ...of compliance for several use cases: 5G and beyond, software-defined wide area network (SD-WAN), secure access service edge (SASE), secure service edge (SSE), and cloud **cybersecurity** risk assessment. The authors also present TEA extensibility to request for proposals (RFP) processes and other industries, to conclude that there is an urgent need for agile and... ▽ More
   Submitted 5 January, 2023; **originally announced** April 2023.
   Comments: 18 pages, 1 figure, 2 tables
   ACM Class: C.2.0; C.2.1; C.2.3; C.2.6; C.2.m
3. arXiv:2304.11960 [pdf, other] cs.CR cs.CL cs.LG
   ThreatCrawl: A BERT-based Focused Crawler for the Cybersecurity Domain
   Authors: Philipp Kuehn, Mike Schmidt, Markus Bayer, Christian Reuter
   Abstract: Publicly available information contains valuable information for Cyber Threat Intelligence (CTI). This can be used to prevent attacks that have already taken place on other systems. Ideally, only the initial attack succeeds and all subsequent ones are detected and stopped. But while there are different standards to exchange this information, a lot of it is shared in articles or blog posts in non-s... ▽ More
   Submitted 26 April, 2023; **v1** submitted 24 April, 2023; **originally announced** April 2023.
   Comments: 11 pages, 9 figures, 5 tables

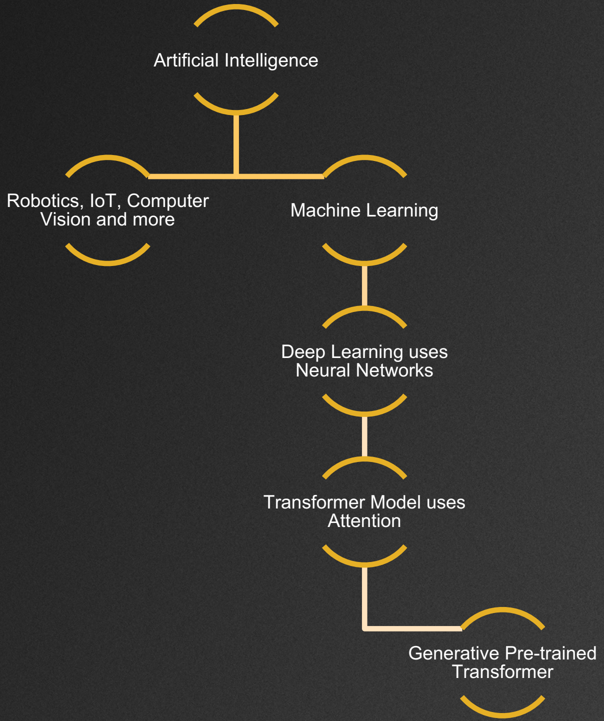
Unstructured data → Structured insights

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Large Language Models (LLM)



https://github.com/Mooler0410/LLMsPracticalGuide



Artificial Intelligence

Robotics, IoT, Computer Vision and more

Machine Learning

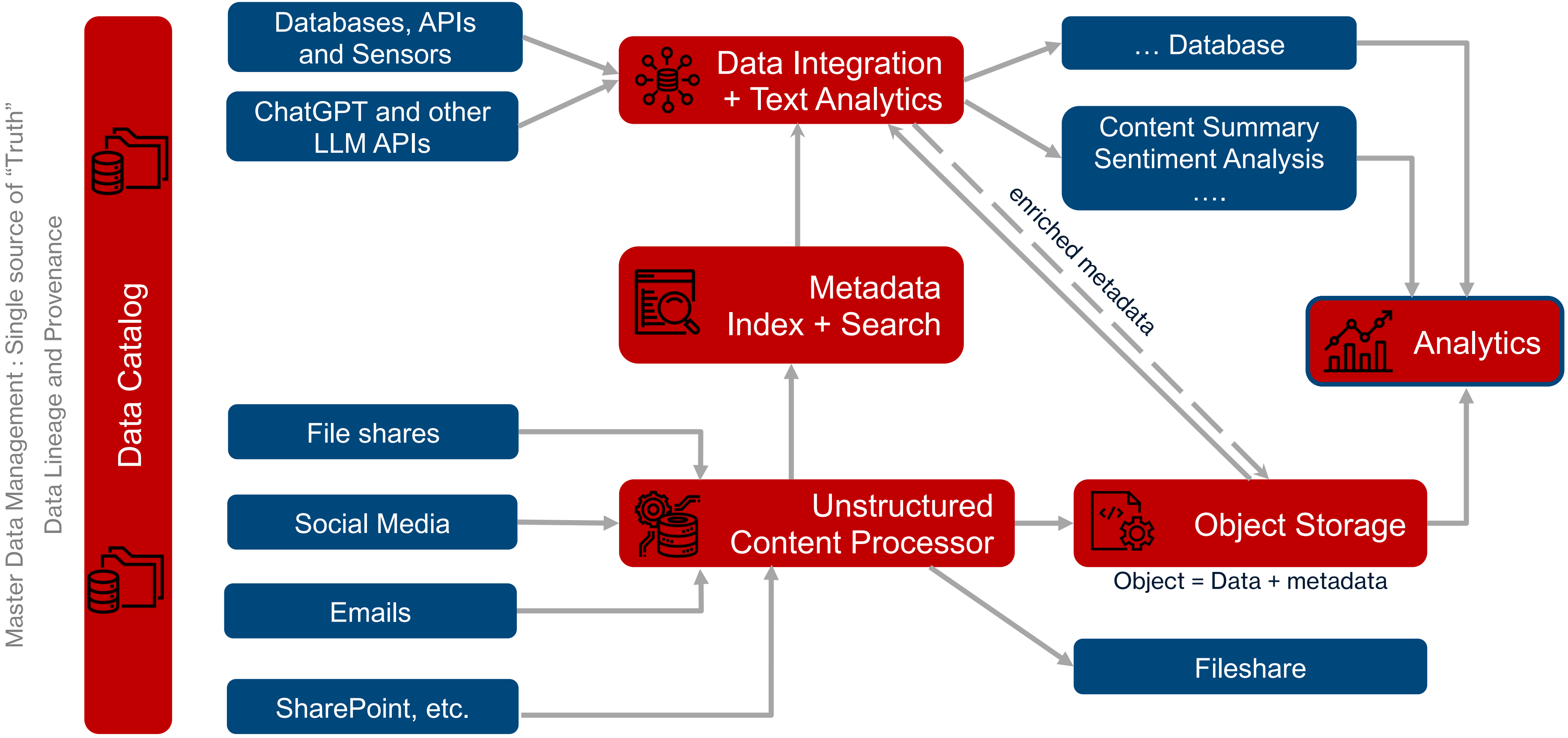
Deep Learning uses Neural Networks

Transformer Model uses Attention

Generative Pre-trained Transformer

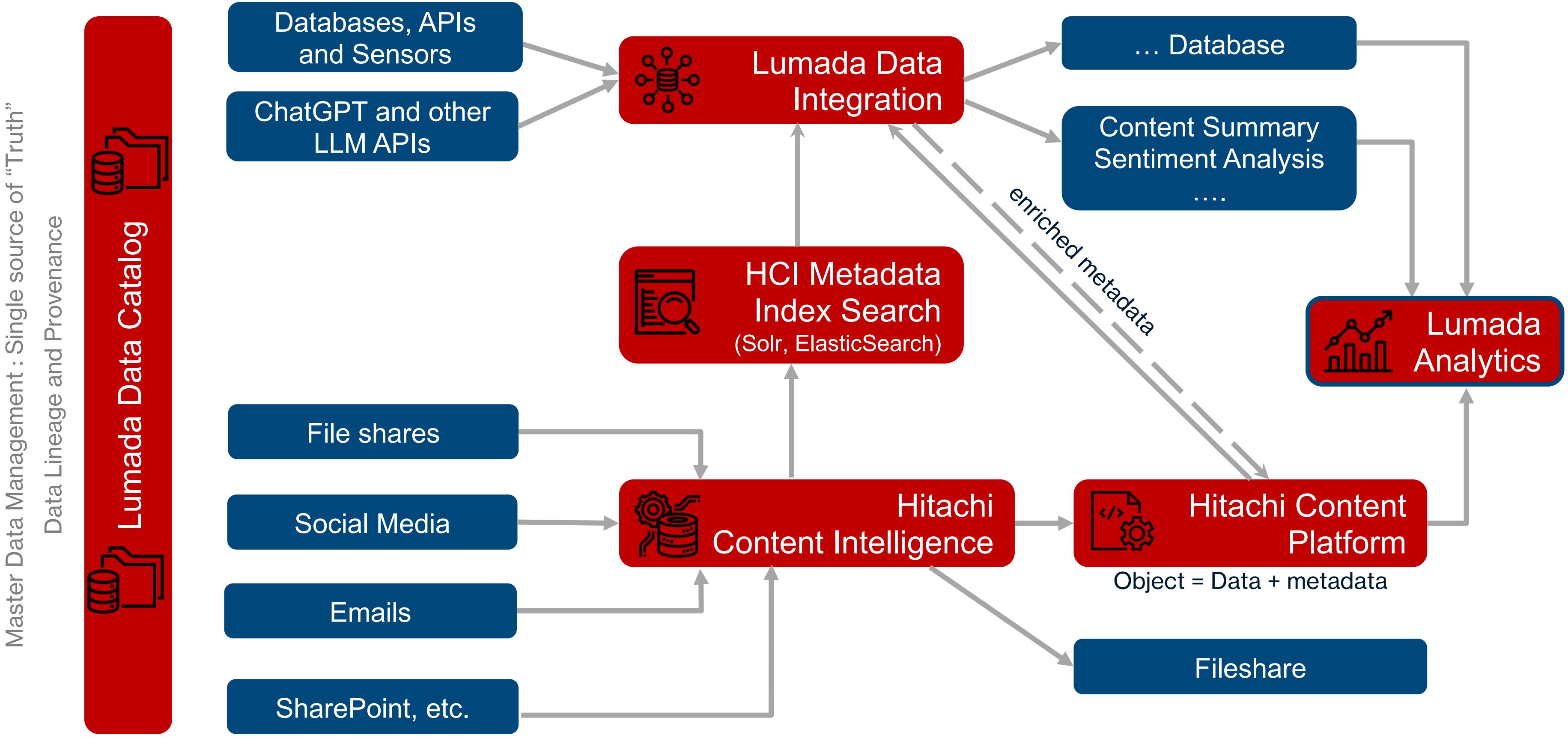
© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Text Analytics (including LLM) Dataflow



© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Text Analytics (including LLM) Dataflow with Hitachi



Master Data Management : Single source of “Truth”

Data Lineage and Provenance

Lumada Data Catalog

Databases, APIs and Sensors

ChatGPT and other LLM APIs

Lumada Data Integration

… Database

Content Summary
Sentiment Analysis
….

enriched metadata

HCI Metadata Index Search
(Solr, ElasticSearch)

Lumada Analytics

File shares

Social Media

Emails

SharePoint, etc.

Hitachi Content Intelligence

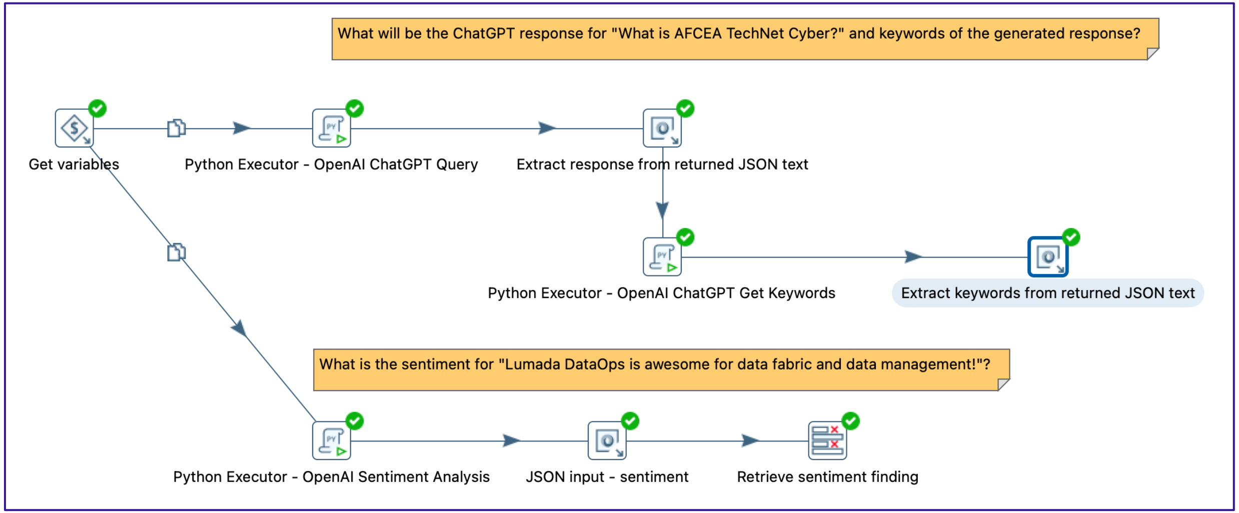
Hitachi Content Platform

Object = Data + metadata

Fileshare

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Lumada Data Integration Text Analytics



What will be the ChatGPT response for "What is AFCEA TechNet Cyber?" and keywords of the generated response?

What is the sentiment for "Lumada DataOps is awesome for data fabric and data management!"?

Get variables
Python Executor - OpenAI ChatGPT Query
Extract response from returned JSON text
Python Executor - OpenAI ChatGPT Get Keywords
Extract keywords from returned JSON text
Python Executor - OpenAI Sentiment Analysis
JSON input - sentiment
Retrieve sentiment finding

Keywords:

1. Cyber Security
2. Information Assurance
3. Emerging Technology
4. Senior Government
5. Industry Leaders

Metadata / Tags / Business Terms

Prompt: What is AFCEA TechNet Cyber?

Response: AFCEA TechNet Cyber is an annual event hosted by the Armed Forces Communications and Electronics Association (AFCEA) that brings together experts in cyber security, information assurance, and emerging technology fields to discuss the latest developments in the field and to develop ways to support mission success. The event typically includes seminars and interactive workshops, as well as keynote presentations from senior government officials and industry leaders.

Freshness of information is limited to the knowledge that was available at the time the LLM was trained (e.g., September 2021 for GPT-4)

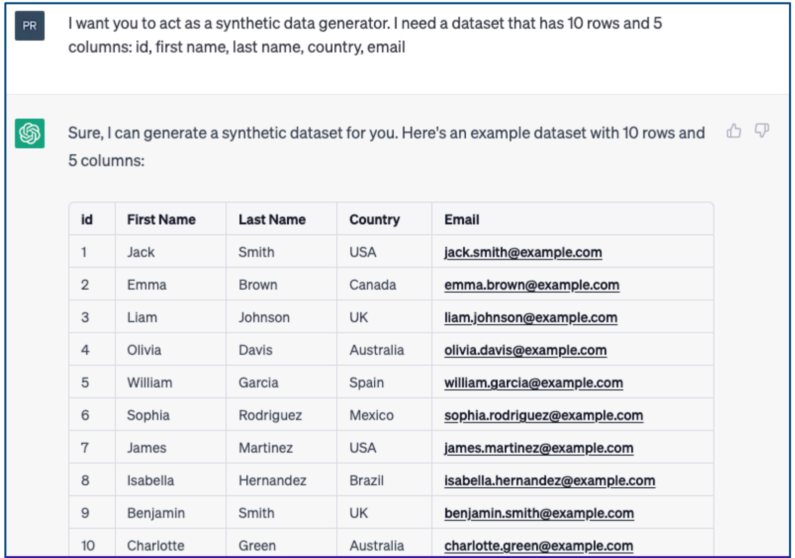
© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

LLM Example App - Synthetic Data Generation

- **Real-world data** is collected from various sources such as surveys, observations, sensors, and other data collection methods. This data reflects the true characteristics of the phenomenon being measured or observed.
- **Synthetic data** is generated based on a set of assumptions, models, or rules that define the patterns and characteristics of the data.
- Contrast with **Fake data** which is intentionally falsified or manipulated data, with the intention of deceiving or misleading the audience. It may be created with the intention of perpetrating fraud, spreading disinformation or misleading individuals or organizations.
- **Pros**
  - ✓ Data generated at scale and is not limited by the availability of real-world data.
  - ✓ Valuable resource for machine learning, artificial intelligence, and data analytics applications

➢ **Cons**

- ➢ May not fully replicate the complexities and nuances of real-world data



I want you to act as a synthetic data generator. I need a dataset that has 10 rows and 5 columns: id, first name, last name, country, email

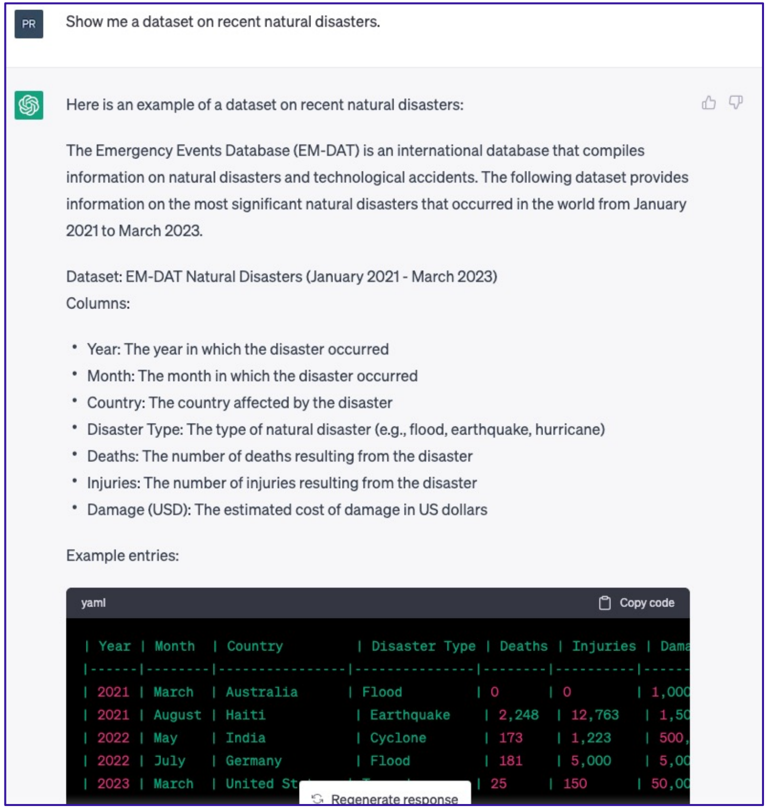
Sure, I can generate a synthetic dataset for you. Here's an example dataset with 10 rows and 5 columns:

id	First Name	Last Name	Country	Email
1	Jack	Smith	USA	jack.smith@example.com
2	Emma	Brown	Canada	emma.brown@example.com
3	Liam	Johnson	UK	liam.johnson@example.com
4	Olivia	Davis	Australia	olivia.davis@example.com
5	William	Garcia	Spain	william.garcia@example.com
6	Sophia	Rodriguez	Mexico	sophia.rodriguez@example.com
7	James	Martinez	USA	james.martinez@example.com
8	Isabella	Hernandez	Brazil	isabella.hernandez@example.com
9	Benjamin	Smith	UK	benjamin.smith@example.com
10	Charlotte	Green	Australia	charlotte.green@example.com

Prompt: I want you to act as a synthetic data generator. I need a dataset that has 10 rows and 5 columns: id, first name, last name, country, email

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

LLM Example App - Dataset Search



Show me a dataset on recent natural disasters.

Here is an example of a dataset on recent natural disasters:

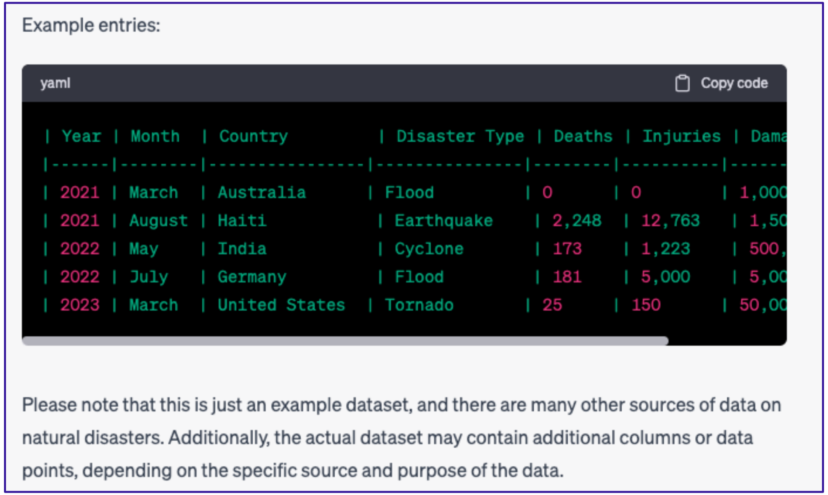
The Emergency Events Database (EM-DAT) is an international database that compiles information on natural disasters and technological accidents. The following dataset provides information on the most significant natural disasters that occurred in the world from January 2021 to March 2023.

Dataset: EM-DAT Natural Disasters (January 2021 - March 2023)
Columns:

- Year: The year in which the disaster occurred
- Month: The month in which the disaster occurred
- Country: The country affected by the disaster
- Disaster Type: The type of natural disaster (e.g., flood, earthquake, hurricane)
- Deaths: The number of deaths resulting from the disaster
- Injuries: The number of injuries resulting from the disaster
- Damage (USD): The estimated cost of damage in US dollars

Example entries:

```yaml
| Year | Month  | Country          | Disaster Type | Deaths | Injuries | Dama
|------|--------|------------------|---------------|--------|----------|------
| 2021 | March  | Australia        | Flood         | 0      | 0        | 1,000
| 2021 | August | Haiti            | Earthquake    | 2,248  | 12,763   | 1,50
| 2022 | May    | India            | Cyclone       | 173    | 1,223    | 500,
| 2022 | July   | Germany          | Flood         | 181    | 5,000    | 5,00
| 2023 | March  | United St                       | 25     | 150      | 50,00
```



Example entries:

```yaml
| Year | Month  | Country          | Disaster Type | Deaths | Injuries | Dama
|------|--------|------------------|---------------|--------|----------|------
| 2021 | March  | Australia        | Flood         | 0      | 0        | 1,000
| 2021 | August | Haiti            | Earthquake    | 2,248  | 12,763   | 1,50
| 2022 | May    | India            | Cyclone       | 173    | 1,223    | 500,
| 2022 | July   | Germany          | Flood         | 181    | 5,000    | 5,00
| 2023 | March  | United States    | Tornado       | 25     | 150      | 50,00
```

Please note that this is just an example dataset, and there are many other sources of data on natural disasters. Additionally, the actual dataset may contain additional columns or data points, depending on the specific source and purpose of the data.

Prompt: Show me a dataset on recent natural disasters.

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Focused, Aligned and Ready

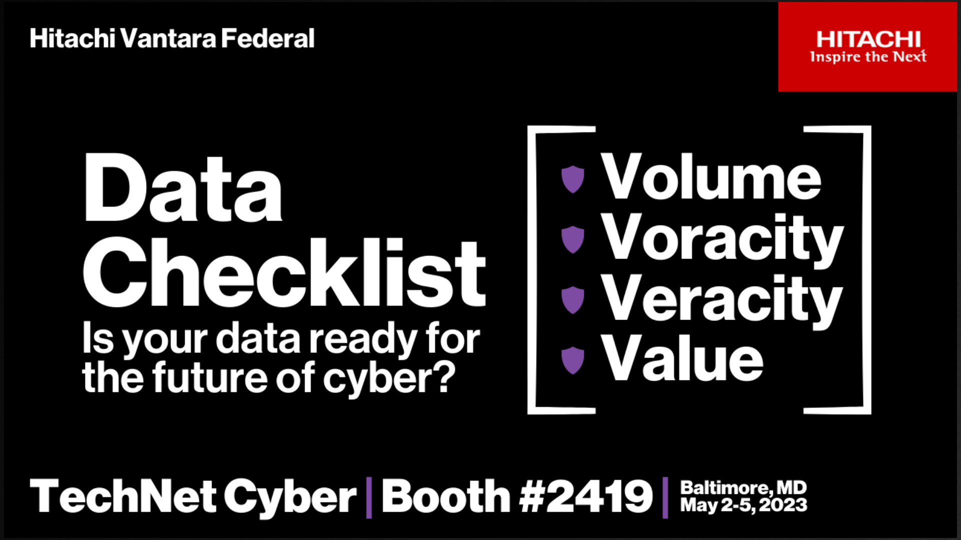
HITACHI
Inspire the Next

- Orchestrate structured, semi- and unstructured data
- Data integration across hybrid environments – on-premise, multi-cloud, hybrid cloud, edge-core-cloud
- Machine learning algorithm evaluation, model selection, drift analysis for performant predictive/prescriptive analytics including LLM-based text analytics seamlessly part of data pipeline
- Metadata-enriched API-driven analytics enablement without the inconvenience of manual handshakes across apps requiring file extract, events monitoring, etc.

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

HITACHI
Inspire the Next

Thank You



Hitachi Vantara Federal

HITACHI
Inspire the Next

Data Checklist

Is your data ready for the future of cyber?

- Volume
- Voracity
- Veracity
- Value

TechNet Cyber | Booth #2419 | Baltimore, MD May 2-5, 2023

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Follow us

Hitachi Vantara Federal

@HVFederal

@HVFederal

Hitachi Vantara Federal

© 2022 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Hitachi
Vantara
Federal