LITERATURE SURVEY – IDEATION PHASE

BY,

TEAM ID : PNT2022TMID08879

S.NO	Paper Name	Author Name	Published Year	Abstract
1	Data analysis and visualization of sales data	Kiran Singh; Rakhi Wajgi	29 February 2016	Data is being generated very rapidly due to increase in information in everyday life. Huge amount of data get accumulated from various organizations that is difficult to analyze and exploit. Data created by an expanding number of sensors in the environment such as traffic cameras and satellites, internet activities on social networking sites, healthcare database, government database, sales data etc., are example of huge data. Processing, analyzing and communicating this data are a challenge. Online shopping websites get flooded with voluminous amount of sales data every day. Analyzing and visualizing this data for information retrieval is a difficult task. Therefore a system is required which will effectively analyze and visualize data. This paper focuses on a system which will visualize sales data which will help users in applying intelligence in business, revenue generation, and decision making, managing business operation and tracking progress of tasks.
2	CityPulse: Large Scale Data Analytics Framework for Smart Cities	Dan Puiu; Payam Barnaghi; Ralf Tönjes; Daniel Kümper; Muhammad Intizar Ali; Alessandra Mileo	2016	Our world and our lives are changing in many ways. Communication, networking, and computing technologies are among the most influential enablers that shape our lives today. Digital data and connected worlds of physical objects, people, and devices are rapidly changing the way we work, travel, socialize, and interact with our surroundings, and they have a profound impact on different domains, such as healthcare, environmental monitoring, urban systems, and control and management applications, among several other areas. Cities currently face an increasing demand for providing services that can have an impact on people's everyday lives. The CityPulse framework supports smart city service creation by means of a distributed system for semantic discovery, data analytics, and interpretation of large-scale (near-)real-time Internet of Things data and social media data streams. To goal is to break away from silo applications and enable cross-domain data integration. The CityPulse framework integrates multimodal, mixed quality, uncertain and incomplete data to create reliable, dependable information and continuously adapts data processing techniques to meet the quality of information requirements from end users. Different than existing solutions that mainly offer unified views of the data, the CityPulse framework is also equipped with powerful data analytics modules that perform intelligent data aggregation, event detection, quality assessment, contextual filtering, and decision support. This paper presents the framework, describes its components, and demonstrates how they interact to support easy development of custom-made applications for citizens. The benefits and the effectiveness of the framework are demonstrated in a use-case scenario implementation presented in this paper.
3	Error data analytics on RSS range-based localization	Shuhui Yang; Zimu Yuan; Wei Li	16 July 2020	The quality of measurement data is critical to the accuracy of both outdoor and indoor localization methods. Due to the inevitable measurement error, the analytics on the error data is critical to evaluate localization methods and to find the effective ones. For indoor localization, Received Signal Strength (RSS) is a convenient and low-cost measurement that has been adopted in many localization approaches. However, using RSS data for localization needs to solve a

				fundamental problem, that is, how accurate are these methods? The reason of the low accuracy of the current RSS-based localization methods is the oversimplified analysis on RSS measurement data. In this proposed work, we adopt a generalized measurement model to find optimal estimators whose estimated error is equal to the Cramér-Rao Lower Bound (CRLB). Through mathematical techniques, the key factors that affect the accuracy of RSS-based localization methods are revealed, and the analytics expression that discloses the proportional relationship between the localization accuracy and these factors is derived. The significance of our discovery has two folds: First, we present a general expression for localization error data analytics, which can explain and predict the accuracy of range-based localization algorithms; second, the further study on the general analytics expression and its minimum can be used to optimize current localization algorithms.
4	A Data Analytics Suite for Exploratory Predictive, and Visual Analysis of Type 2 Diabetes	Nada Y. Philip; Manzoor Razaak; John Chang; Suchetha. M; Maurice O'Kane;	27 January 2022	Long-term management of chronic disorders such as Type 2 Diabetes (T2D) requires personalised care for patients due to variation in patient characteristics and their response to a specific line of treatment. The availability of large volumes of electronic records of T2D patient data provides opportunities for application of big data analysis to gain insights into the disease manifestation and its impact on patients. Data science in healthcare has the potential to identify hidden knowledge from the database, re- confirm existing knowledge, and aid in personalising treatment. In this paper, we present a suite of data analytics for T2D disease management that allows clinicians and researchers to identify associations between different patient biological markers and T2D related complications. The analytics suite consists of exploratory, predictive, and visual analytics with capabilities including multi-tier classification of T2D patient profiles that associate them to specific conditions, T2D related complication risk prediction, and prediction of patient response to a particular line of treatment. The analytics presented in this paper explore advanced data analysis techniques, which are potential tools for clinicians in decision-making that can contribute to better management of T2D.
5	Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data	Mingchen Feng; Jiangbin Zheng; Jinchang Ren; Amir Hussain; Xiuxiu Li; Yue Xi;	22 July 2019	Big data analytics (BDA) is a systematic approach for analyzing and identifying different patterns, relations, and trends within a large volume of data. In this paper, we apply BDA to criminal data where exploratory data analysis is conducted for visualization and trends prediction. Several the state-of-the-art data mining and deep learning techniques are used. Following statistical analysis and visualization, some interesting facts and patterns are discovered from criminal data in San Francisco, Chicago, and Philadelphia. The predictive results show that the Prophet model and Keras stateful LSTM perform better than neural network models, where the optimal size of the training data is found to be three years. These promising outcomes will benefit for police departments and law enforcement organizations to better understand crime issues and provide insights that will enable them to track activities, predict the likelihood of incidents, effectively deploy resources and optimize the decision making process.
6	A Generic Data Analytics System for Manufacturing Production	Hao Zhang, Hongzhi Wang , Jianzhong Li, and Hong Gao	2 June 2018	The increase in the amount of manufacturing information available means that big data can be collected and, with appropriate deep analysis, could be of great value to manufacturers. However, most small manufacturers cannot afford the overhead of a professional data analytics team. To address this problem, in this paper a generic data analytics system, Generic Manufacturing Data Analytics system (GMDA), is proposed. This system can perform most manufacturing

				data analytics tasks and users can easily carry out data analysis even if they have no prior knowledge or experience of data analytics. To establish such a system, we designed an abstract language, GMDL, to describe the manufacturing data analytics tasks. Aimed at factory data analytics, several algorithms were selected, tuned, optimized, and finally integrated into the system. Some noteworthy techniques were developed in GMDA such as proper algorithm selection strategy and an optimal parameter determination algorithm. Case studies show the practicability and reliability of the system.
7	Distributed Data Strategies to Support Large-Scale Data Analysis Across Geo-Distributed Data Centers	TAMER Z. EMARA AND JOSHUA ZHEXUE HUANG	September 29, 2020	As the volume of data grows rapidly, storing big data in a single data center is no longer feasible. Hence, companies have developed two scenarios to store their big data in multiple data centers. In the first scenario, the company's big data are distributed in multiple data centers without data replication. In the second scenario, data are also stored in multiple data centers but important data are replicated in these data centers to increase data safety and availability. However, in these scenarios, analyzing big data distributed in multiple data centers becomes a challenging task. In this paper, we propose two data distribution strategies to support big data analysis across geo-distributed data centers. In these strategies, we use the recent Random Sample Partition data model to convert big data into sets of random sample data blocks and distribute these data blocks into multiple data centers either without replication or with replication. In analyzing big data in multiple data centers without replication, we randomly select samples of data blocks from multiple data centers and download the sample data blocks to one data center for analysis. In the second strategy with replication of data blocks, we can analyze big data on any data center by randomly selecting a sample of data blocks replicated from other data centers. This strategy avoids data transformation between data centers. We demonstrate the performance of the two strategies in big data analysis by using simulation results produced on one local data center and four AWS data centers in North America, Asia, and Australia.
8	Migration-Based Online CPSCN Big Data Analysis in Data Centers	XIN LI, LIANGYUAN WANG1 , ZHEN LIAN	February 28, 2018	It is critical to schedule online data-intensive jobs effectively for various applications, including cyber-physical-system and social network system. It is also useful to support timely decision making and better prediction. In this paper, we investigate the online job scheduling problem with data migration for global job execution time reduction. We first establish a time model based on the real experimental results, and propose an online job placement algorithm by taking into account the benefit of both instantaneity and locality for the jobs. We then introduce data migration to the job placement algorithm. The core idea is to make a tradeoff between the migration cost and remote access cost. The simulation results demonstrate that our algorithm has a significant improvement than FIFO, and data migration shows effectiveness on global job execution time reduction. Our algorithms also provide an acceptable fairness for jobs.
9	Agriculture Data Analytics in Crop Yield Estimation: A Critical Review	Sagar Bm	2018	The use of technology in agriculture has increased in recent year and data analytics is one such trend that has penetrated into the agriculture field. The presentstudy givesinsights onvarious data analytics methods appliedto cropyield predictionand also signifies the important points in the proposed area of research.
10	The Impact of Data Analytics in Crop	Swarupa Rani A	2017	Discussed theapplication ofmathematical modellike fuzzy logic designs in optimizationof thecrop yield, artificial neural networks in validation studies, geneticalgorithms designs in accessing the fitness of the model applied, decision trees, andsupport vectormachines to

	Management based on Weather Conditions			study soil, climate conditions and water regimes related to crop growth and pest management in agriculture.
11	A Study on Crop Yield Forecasting Using Classification Techniques	R.Sujatha, Dr.P.Isakki Devi	2016	Discuss the importanceof comparing previous agricultural data with present to identify optimum condition favor enhanced crop yield. Envisaged the importance ofbest crop selection depending on the season and the climatic factors which supports enhanced crop yield.
12	Prediction of Crop Yield using Regression Analysis	V. Sellamand E.Poovammal	2016	Regression analysis was carried out to find the relationship among the parameters i.e Area under Cultivation (AUC), Annual Rainfall (AR) and Food Price Index (FPI) which influences the final crop yield and reported that the crop yield principally depends on the Annual Rainfall (AR).
13	Data requirements for reliable crop yield simulations and yield-gap analysis	Patricio Grassinia, Lenny G.J. van Bussel, Justin Van Warta, Joost Wolf, Lieven Claessens, d, Haishun Yanga, Hendrik Boogaarde, Hugo de Groote,Martin K. van Ittersumb, Kenneth G. Cassman	2015	Presented acase study (Nebraska - USA and ata nationalscale for Argentina and Kenya) on the application of an explicit rationale design approach in identifying the data sources which simulates Crop (maize) yield and also helps in quantifying the maize yield gaps. Suggested the robust guidelinesfor analyzingthe cropyield gaps, accessing the climate and land use changes at global level to address the issues of crop yield.
14	A Survey on Crop Yield Prediction based on Agricultural Data	Dhivya B H, Manjula R, Siva Bharathi S, Madhumathi R	2017	Presented asurvey onthe differentalgorithms appliedin the assessment and prediction of crop yield. Discussed about the mechanismof knowledgethe discovery in Agricultural data estimation
15	The use of satellite data for crop yield gap analysis	David B. Lobell	2013	Discussed theuse ofremote sensingtechnology toidentify andmeasure thecauses of yield gaps andthe assess the impact on the overall crop yield. Reported very simplemethodologies tomeasure theyield difference with respect to season, environment and the land use.