My main research interests are in the fields of database systems and big data management. My research specifically focuses on the design and evaluation of *big data stream processing* algorithms in distributed Cloud and Edge computing settings. Big data streams from dynamic application scenarios are served unceasingly to data stream processing (DSP) systems, typically coming from sensor measurements, mobility trackers and other IoT devices, in addition to microblogging websites. Such data normally has three associated dimensions: (1) geolocation, describing the geography where it has been collected, (2) time, indicating the time it has been generated, and (3) context, in the form of additional information characterising the environment surrounding the measurement, such as weather conditions (temperature, humidity, etc.,). The pivotal research question in such scenarios is *how to manage and process the multidimensional big data with QoS guarantees*. This question is broad, and under its umbrella, it incorporates various aspects that collectively help characterising the theoretical design and system development of reliable distributed systems. Most systems in the relevant state-of-art (1) do not feature QoS guarantees as a first-class citizen in their intrinsic designs and (2) do not consider a combination between at least two of the three dimensions (geolocation, time, and context) collectively. Having said that, the general theme characterizing my research is to design, implement and evaluate new algorithms and tools for the efficient processing and management of multidimensional big data in dynamic application scenarios that are prevalent in smart cities and urban informatics.

My research is elegantly inspired by real-world application scenarios in smart cities and urban planning. I always focus on system and theoretical aspects for developing sustainable and highly performing big data management tools that work on real data originating in such scenarios. During my research journey, I have had several gratifying and rewarding opportunities to collaborate closely with several researchers and domain experts in various fields pertinent to big data management, machine learning, Cloud/Edge computing and deep learning. This combination shaped the synergy of my previous, current, and perspective research interests and directed me to recognize the applicability and impact my research can have on our world.

Research During my PhD

During my PhD at University of Bologna (2016 – 2020), I focused on designing a novel system architecture for geospatial Cloud-based big data management with QoS. I focused on answering the following research questions: *(1) how to efficiently store multidimensional data in distributed scalable storage systems, (2) how to process batch and stream big multidimensional data, at scale, with QoS guarantees, (3) how to maintain stability of distributed systems for multidimensional big data processing under fluctuations in data arrival rates and shape.*
Unlike traditional management of big data streams where QoS guarantees are neglected, I contemplated an architecture with intrinsically incorporated QoS awareness, and designed spatial data stream management system (SpatialDSMS), which encompasses QoS proprietary utilities for the scalable storage and stream processing of multidimensional data in distributed systems. Under this system architecture, I developed several big multidimensional management tools, which are explained in three main branches.

Scalable storage. Traditional distributed storage assumes geospatial data is normally distributed and stores geolocated data in far apart distributed storage nodes. However, several real-world applications (sensor measurements, vehicle mobility, geotagged microblogs) generate highly skewed data, which need to be stored in nearby nodes of a distributed computing cluster to allow seamless targeted query operations. To close this gap, I designed a custom geospatial data partitioning algorithm, GSS [1] that incrementally splits the input georeferenced data stream in a way that guarantees a plausible trade-off between load balancing and geospatial data colocation, sending geometrically-nearby data to same nodes.
The applicability of GSS in dynamic application scenarios is significant. I applied GSS to urban planning [1]. Specifically, I developed the SpatialNoSQL [1] system, which exploits GSS for advanced analytics of mobility data at scale. The contributions are (1) *advanced urban planning analytics*: it successfully employs novel two-level indexing for efficient geospatial queries that incorporates spatial join predicates, operating

faster as opposed to traditional methods, and (2) *adaptive data reduction*: it elegantly reduces data volumes by means of geospatial aggregations, showing the trends of how mobility in a city is changing across time.

In the same vein, I designed SCAP [2], a spatial col-locality aware partitioner for efficiently processing massive amounts of georeferenced data within in-memory DSP systems, at scale, with QoS guarantees. I proved the successful applicability of SCAP to dynamic urban planning by exploiting its characteristics in a DSP system that I designed, SpatialBPE [2]. In a more concise sense, SpatialBPE employs SCAP with additional query optimizers for mobility data clustering scenarios.

SpatialNoSQL (and GSS) SpatialBPE (and SCAP) are extensions of our previous works [3-5] which we have presented in top-tier venues, where we have received rave feedback from industry practitioners and research community.

Stream. Another integral component of my PhD thesis is on data stream processing, where georeferenced data arrives in a fast pace that exceeds capacities of DSP systems. The goal is enabling DSP systems to withstand massive data loads. To achieve this, I designed spatial aware online sampling (SAOS), an efficient algorithm for withdrawing geospatially representative data samples on-the-fly from a data stream [6].
Georeferenced data streams are highly skewed and multidimensional, however, methods in relevant literature are unaware of those characteristics. Consequently, sampling results are unrepresentative, which leads to inaccurate analytical results. Based on that, I designed SAOS, which achieves orders of magnitude improvement in terms of accuracy and speed as opposed to state-of-art counterparts. I applied SAOS in urban planning context with massive real-world mobility data, which attests to its applicability to dynamic application scenarios. I also designed ex-SAOS [7], which is an extended version of SAOS that captures representative geospatial data stream samples on a coarser level with more accuracy.

The overarching traits of SAOS helped me win Microsoft ‘AI for Earth’ grant for two years in row (July 2020 – July 2022).

Multidimensional join. Numerous are the application scenarios that require more advanced analytics by instantly joining fast-arriving georeferenced big data streams with a disk-resident counterpart. For example, parametrized mobility data (i.e., geolocation information represented as longitude/latitude pairs) need to be joined to shape files (i.e., representing the geography where mobility traces belong) for insightful analytics. Joining parametrized geospatial data on-the-fly is computationally expensive and can easily bring a DSP system into halt during peak hours of arrival rates. I contributed significantly to this direction by introducing novel adaptive algorithm for stream-static geospatial join processing.
Specifically, I designed SpatialSSJP [8] for efficiently joining georeferenced big data streams with disk-resident data, at scale, with QoS guarantees. I applied SpatialSSJP to dynamic urban planning scenarios for efficient massive mobility data analytics.

Current Research

The approach I take as a postdoctoral researcher is to expand my previous research gradually in related directions. I aim at building up my research expertise with components that can have a real impact in our world, which will also help me in building fruitful collaboration with diverse distinguished researchers who are working in various pertinent domains. I started considering the two other equally important dimensions in multidimensional data coming from dynamic application scenarios: time and context.

Approximate query processing. I have recently designed ApproxSSPS [9], a novel spatial data stream processing system that can deliver real accurate results in a timely manner, by dynamically specifying the limits on data samples. ApproxSSPS features a controller that interactively learns latency statistics and calculates proper sampling rates to meet latency or/and accuracy targets. An overarching trait of ApproxSSPS is its ability to strike a plausible balance between latency and accuracy targets.

Multidomain processing. To further push my research agenda to bordering fields, I decided to investigate the challenges associated with management and processing of heterogeneous big multidimensional data coming from various domains. In a more precise sense, I recently (2021) designed a novel multidimensional join method for joining georeferenced mobility data with georeferenced time-series meteorological measurements representing indications of pollution (particulate matters, PM10 and PM2.5) in metropolitan cities. Specifically, I designed and evaluated MeteoMobil [10] for the combined analytics of integrated information representing mobility and environment conditions. MeteoMobil can efficiently tag mobility data with environment (e.g., meteorological) data, at scale. Also, it features an SQL-alike API for simplifying queries such as aggregation, grouping and statistics on environmentally-tagged mobility data. In this case, additional weather information from the meteorological domain is treated as a contextual data that enriches related mobility data and broadens the exploitability of such data in an unprecedented manner. Stated another way, it enables performing advanced analytics pertinent to dynamic smart city application scenarios that are unachievable otherwise. For example, informing the decisions for better urban planning that help in reducing the adverse effects of particulate matters on human health. Based on this new system, I was invited by Microsoft to give a talk in FOSS4G 2021, the largest global gathering for geospatial software.

MeteoMobil [10] and ApproxSSPS [9], in addition to other forthcoming systems are part of a Microsoft 'AI for Earth' grant that I won for two consecutive years (July 2020 – July 2022).

Context-aware recommender systems. Although I started part of this work during my PhD while spending three months as a visiting researcher, it came to fruition only the moment I started my current position as a postdoctoral fellow. Having focused mainly on geospatial dimension during my PhD. In late 2019, I realized the importance of bringing the two other dimensions (time and context) into the stage. Contextual information in dynamic application scenarios is equally important and can enhance the decision-making process in urban planning. I recently focused on the incorporation of contextual information in deep learning-based recommender systems. specifically, I designed CA-NCF [11], which is a first-in-class context-aware deep learning-based recommender system. I have applied it to various scenarios, including movie and trip recommendations. However, my intention in starting this research line is to soon design methods that tag georeferenced mobility data with additional contextual information from other domains such as meteorological and climatology data. Thereafter, I plan to design machine learning and deep learning models for learning trends from the conjunction that such multidomain data brings. For example, designing a recommendation system that recommends trip routes for lightweight dwellers in polluted metropolitan cities in such a way that helps them avoid the adverse effects that polluted routes may impose on their health. I further recently designed US-NCF [12], which employs a new contextual incorporation method for deep learning based recommender systems. It considers situations where social interactions data from social websites need to be gathered to tag recommendation data coming from other domains such as movie recommendation websites.

CA-NCF proposal helped me win a merit-based grant for the 'mobility of young researchers, Marco Polo Program' funded by the University of Bologna, for a research visit to universidad de Extremadura in Spain for three months.

Based on my works in big geospatial data management, I recently proposed to design and teach a unique PhD course in the Department of Computer and Engineering at University of Bologna. My proposal was accepted by the department council. The course is unique in the sense it is, to the best of my knowledge, the first of its kind that covers advanced aspects of designing distributed geospatial data-intensive systems.

Research Plans

I am totally thrilled by the opportunities my research theme offers. I plan to expand my research areas in several equally important and complementing directions, always pertinent to the general theme of big data management and databases systems.

Scalable processing of georeferenced time-series data. The focus so far has been geared toward the geospatial dimension in dynamic application scenarios. The time dimension is, however, equally important. Numerous scenarios in smart cities and urban planning require specifically to filter geo-referenced time series data on both dimensions. Georeferenced time series data are those records that are associated with locational tags. For example, to analyse trends of vehicle mobility, a 'moving average' of the vehicle's speed need to be computed continuously as time ticks forward. This is equivalent to calculating the average in each region of a city across time. Thus, two dimensions are involved, geolocation and time. To my knowledge, current works focus on optimizing each dimension independently. In my future research I aim at extending my previous and current works on big geospatial data management by involving the time aspect intrinsically. This will include designing new algorithms and methods for indexing georeferenced time-series data, in addition to custom methods for join processing and partitioning in distributed computing systems, considering both dimensions collectively, time and location.

Offloading part of the processing load to Edge devices. In my previous research, I focused specifically on designing methods to operate in distributed computing systems that are deployed in Cloud settings. Nevertheless, sending all data stream to a Cloud deployment is not always necessary. Other research issues I am tackling are the following. What are some of the best methods and architectural models to meet performance requirements in dynamic application scenarios with fluctuating performance parameters (data arrival and operator service rates)? which architectural model should be selected for a QoS-aware data management of georeferenced time-series data: Cloud-based or Edge-Cloud? Adding IoT Edge devices to the equation is another issue and has its complexities that normally do not affect Cloud-solo architectures. For example, in my previous research, I have designed SAOS [6] for capturing geospatially-representative samples from fast arriving data streams during peak hours and answering geolocation queries with rigorous error-bounds. I plan soon to tweak this method and offload it to Edge devices so that representative samples are selected in Edge devices instead of Cloud, thus moving less data to the Cloud for further processing and relieving the stress on communication network. I have a firm belief that my expertise in big multidimensional data stream processing gives me an advantage for solving such problem.

Data stream visualization and anomaly detection. I plan in the near term to cover two major aspects of processing multidimensional data streams: *visualization* and *anomaly detection*.
Multidimensional big data streams typically fluctuate in arrival rates and skewness, causing sometimes what is known as concept drift. Massive amounts of data arrive momentously that need to be visualized for better-informed insightful analytics. The size and diversity of data hinders the application of appropriate distributed visualization techniques and could bring them easily into halt. I plan to investigate this link of research to design and build new methods and systems for improving state-of-art visualization tools for multidimensional data streams, considering the three dimensions: location, time, and context.
In a related context, anomalous data provides more insights than typical same-pattern counterparts. When it comes to multidimensional data stream processing, anomaly detection is becoming more challenging given massive size and skewness of data shapes. I plan to explore this direction and design new algorithms for anomaly detection in big multidimensional data streams.

Big data management for studying urban traffic emissions. In environmental monitoring, model-driven approaches (e.g., COPERT, MOVES) compute urban macro-emissions, but fail to provide traffic emission prediction on granular levels (e.g., street level). As a way of contrast, data-driven approaches integrate vehicle mobility GPS and road network data with weather and meteorological data to study granular vehicle emission trends. They employ deep neural network (DNN) models by applying map-matching (e.g., trajectory-to-location join) algorithms to estimate the spatial distributions of GPS trajectories data in urban regions. The resulting views can reflect the traffic status such as vehicle population, mileage, speed in, for example, a grid-like format. Those geo-statistics can then feed NN for multiple purposes. For example, generating predictions of those statistics in the future, emission station location prediction, predicting vehicles traffic emissions for each region of a city, studying the evolution of traffic emissions. However, those map-matching and geospatial join algorithms are extraordinarily expensive, particularly given the

avalanches of fast-arriving multidimensional mobility data streams. I plan soon to extend my current research so that it covers, more intensively, those aspects. This includes designing novel methods for efficient geospatial joins (such as trajectory-to-location join and trajectory similarity joins), also taking the time dimension into consideration. It will also include designing novel data stream spatiotemporal and time-series indexing methods dedicated for this domain.

QoS-aware mobility digital twins. The computational and storage capacities of IoT devices are parsimonious and can be easily depleted by the overwhelming flow of geo-referenced data streams. Porting those workloads to powerful Cloud deployments is then incumbent. I aim at expanding my current research in a direction that targets building a novel architecture at the confluence between IoT sources, Edge, and Cloud digital twins. Stated another way, I envision a QoS-aware architecture that constitutes three layers: physical layer, communication layer and Cloud layer. Huge amounts of geo-referenced data streams generated in the physical layer will be sampled, pre-processed, transformed and loaded to corresponding digital twin replicas in the Cloud layer. The Cloud deployment is then responsible for the costly storage and computation of such multidimensional data, generating insights that is to be fed to the actuators in the physical layer for interactive and proactive actions. Computations include applying machine learning and deep learning models, which normally can not be applied in the Edge near the IoT sources. Data includes human and vehicle mobility data, in addition to traffic, map, meteorological and weather data. Due to the multidimensionality of those datasets, I aim at designing novel methods for interfacing those layers with QoS guarantees, specifically for integrating data from multiple IoT sources of the physical space into the digital twin replicas of the Cloud, taking into considerations the three data indispensable dimensions (geospatial, temporal, and contextual). A case scenario is the following, a data-driven vehicle emission calculator connected to a network of Cloud-hosted vehicle digital twins can be used to calculate, on-the-fly, fine-grained street-level emissions, thereafter those emission levels can be published to mobile-based maps of city dwellers who wish to avoid polluted streets while jogging within the city.

From an applicability and system perspective, I plan to establish a rigorous team of researchers (research assistants, undergraduate and graduate researchers) and work in close collaboration with industry practitioners to design and develop new systems for big multidimensional data management covering the three dimensions: geolocation, time, and context. Promising domains include (1) climate change analytics and (2) urban planning. I plan to openly release our system artifacts to a broad research and industry community, thus bringing my research ideas much closer toward the technology transformation edge.

Other Research Works

I also worked intermittently with data warehousing and data lakes design. During my involvement in a funded project known as SACHER [13] project, I designed SACHER MuSE CH [13], a novel NoSQL-based data lake system for cultural heritage data management. Prior to my PhD studies, I have a significant work in data warehousing while I was a Lecturer at University of Business and Technology, where I designed a new data warehouse system for decision support in higher education [14]. At the time, I received a fund for my proposal from University of Business and Technology, for an amount that almost is equal to $8K.

References

[1] I. M. Al Jawarneh, P. Bellavista, A. Corradi, L. Foschini, and R. Montanari, "Efficient QoS-Aware Spatial Join Processing for Scalable NoSQL Storage Frameworks," *IEEE Transactions on Network and Service Management,* 2020.

[2] I. M. Al Jawarneh, P. Bellavista, A. Corradi, L. Foschini, and R. Montanari, "Locality-Preserving Spatial Partitioning for Geo Big Data Analytics in Main Memory Frameworks," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*, 2020: IEEE, pp. 1-6.

[3] I. M. Al Jawarneh, P. Bellavista, F. Casimiro, A. Corradi, and L. Foschini, "Cost-effective strategies for provisioning NoSQL storage services in support for industry 4.0," in *2018 IEEE Symposium on Computers and Communications (ISCC)*, 2018: IEEE, pp. 01227-01232.

[4] I. M. Aljawarneh, P. Bellavista, A. Corradi, R. Montanari, L. Foschini, and A. Zanotti, "Efficient spark-based framework for big geospatial data query processing and analysis," in *2017 IEEE Symposium on Computers and Communications (ISCC)*, 2017: IEEE, pp. 851-856.

[5] I. M. Al Jawarneh, P. Bellavista, A. Corradi, L. Foschini, R. Montanari, and A. Zanotti, "In-memory spatial-aware framework for processing proximity-alike queries in big spatial data," in *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2018: IEEE, pp. 1-6.

[6] I. M. Al Jawarneh, P. Bellavista, L. Foschini, and R. Montanari, "Spatial-Aware Approximate Big Data Stream Processing," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019: IEEE, pp. 1-6.

[7] I. M. Al Jawarneh, P. Bellavista, A. Corradi, L. Foschini, and R. Montanari, "Spatially Representative Online Big Data Sampling for Smart Cities," in *2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2020: IEEE, pp. 1-6.

[8] A. Jawarneh and I. M. Hasan, "Quality of service aware data stream processing for highly dynamic and scalable applications," 2020.

[9] I. M. Al Jawarneh, P. Bellavista, A. Corradi, L. Foschini, and R. Montanari, "QoS-Aware Approximate Query Processing for Smart Cities Spatial Data Streams," *Sensors,* vol. 21, no. 12, p. 4160, 2021.

[10] I. M. Al Jawarneh, P. Bellavista, A. Corradi, L. Foschini, and R. Montanari, "Efficiently Integrating Mobility and Environment Data for Climate Change Analytics," in *2021 IEEE 26th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2021: IEEE, pp. 1-5.

[11] I. M. Al Jawarneh *et al.*, "A pre-filtering approach for incorporating contextual information into deep learning based recommender systems," *IEEE Access,* vol. 8, pp. 40485-40498, 2020.

[12] I. M. Al Jawarneh, P. Bellavista, A. Corradi, L. Foschini, and R. Montanari, "Context Incorporation Techniques for Social Recommender Systems," in *ICC 2021-IEEE International Conference on Communications*, 2021: IEEE, pp. 1-6.

[13] S. Bertacchi *et al.*, "SACHER Project: A cloud platform and integrated services for cultural heritage and for restoration," in *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*, 2018, pp. 283-288.

[14] I. M. Aljawarneh, "Design of a data warehouse model for decision support at higher education: A case study," *Information Development,* vol. 32, no. 5, pp. 1691-1706, 2016.