PAIRS: A scalable geo-spatial data analytics platform

Levente J Klein*, Fernando J Marianno, Conrad M Albrecht, Marcus Freitag, Siyuan Lu, Nigel Hinds, Xiaoyan Shao, Sergio Bermudez Rodriguez[1], Hendrik F Hamann
IBM TJ Watson Research Center Yorktown Heights, NY 10598
[1]Osram Sylvania, Beverly, MA 01915
* Email: kleinl@us.ibm.com

Abstract—**Geospatial data volume exceeds hundreds of Petabytes and is increasing exponentially mainly driven by images/videos/data generated by mobile devices and high resolution imaging systems. Fast data discovery on historical archives and/or real time datasets is currently limited by various data formats that have different projections and spatial resolution, requiring extensive data processing before analytics can be carried out. A new platform called Physical Analytics Integrated Repository and Services (PAIRS) is presented that enables rapid data discovery by automatically updating, joining, and homogenizing data layers in space and time. Built on top of open source big data software, PAIRS manages automatic data download, data curation, and scalable storage while being simultaneously a computational platform for running physical and statistical models on the curated datasets. By addressing data curation before data being uploaded to the platform, multi-layer queries and filtering can be performed in real time. In addition, PAIRS offers a foundation for developing custom analytics. Towards that end we present two examples with models which are running operationally: (1) high resolution evapo-transpiration and vegetation monitoring for agriculture and (2) hyperlocal weather forecasting driven by machine learning for renewable energy forecasting.**

Keywords: big data analytics; GIS; Hadoop & HBase for geo-spatial data; MapReduce; data management systems; machine learning

I. INTRODUCTION

Digitization of the "world" is changing many industries including the way in which geospatial data is analyzed. With daily imaging of earth surface by multiple satellites, spatial and temporal correlations can be established between locations and events in real time. In addition to satellite images, weather or climate models are updated multiple time per days generating insight into the atmosphere–earth interaction and its impact on environment, business activity, and human life. While static or reanalysis studies were carried out, in the past, for example to understand deforestation [1], land use [2], or urban area expansion [3], it is expected that, in the future, these models will run in real time.

The exploding volume of geospatial data requires the development of a scalable platform that can fuse multiple data layers and combine them with local measurements from mobile devices or sensor networks. Such a platform should not be only a data repository but should also serve as modeling and analytics platform [4]. Combination of data and analytics can be used for running global models like crop production, water availability, soil moisture or urban expansion. These models can be updated in real time using the latest available datasets and provide insight into dynamic changes like flooding, wildfires, or landslides as they develop.

The daily generation rate for selected satellite and weather/climate data sets is summarized in Fig. 1. More than 700 Landsat 8 [5] tiles are acquired daily in addition to 400 Landsat 7 tiles; this generates in excess of 1 Terabyte of geospatial data per day. Similarly, Moderate Resolution Imaging Spectroradiometer (MODIS) [6] instrument data generation rate approaches 1 Terabyte/day, acquiring data in 250 spectral bands. Based on this data, new data products are being derived to analyze earth's land, ocean, and atmosphere generating even more data. By far the largest geospatial data volume is generated by numerical weather and climate forecasting such as Global Forecast System (GFS), Global Ensemble Forecast System (GEFS), Climate Forecast System (CFS) in the US [7], and the European Centre for Medium-Range Weather Forecasts (ECMWF) model. Weather and climate models generated by ECMWF are in excess of 12 Terabyte/day [8].

If weather and satellite data are to be analyzed and integrated into models before the forecasted data becomes obsolete, then data processing should be accelerated through parallelization. One way to achieve this would be to have all data layers curated and homogenized before being uploaded to the platform, eliminating the time required for data preprocessing. The data curation require data validation, verification, and alignments spatially and temporarily, such that these layers are ready to be integrated into physical and statistical models without the need for data download, validation, and preprocessing.

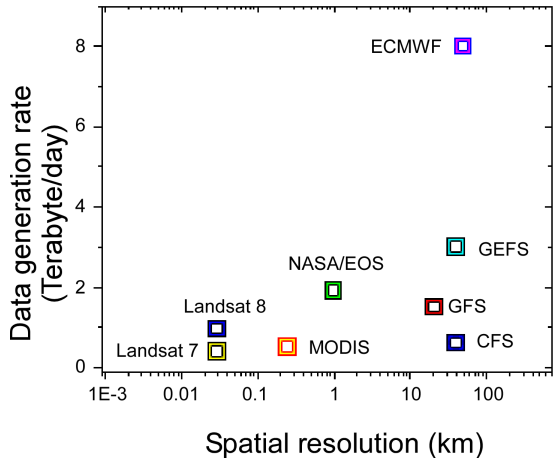

Figure 1. Geo-spatial data generation rate for weather and satellite data. Except for the European Centre for Medium-Range Weather Forecasts (ECMWF), data are published at rates of ~1 Terabyte per day - independent of the spatial resolution of the raster data.

Satellite and weather have well-defined data format and associated metadata, requiring less custom effort for processing. A more challenging task is unstructured data processing. In the unstructured data category falls data coming from mobile devices, social networks, edge devices, or sensor networks that are less likely to have similar format or accuracy than the structured data.

There are more than 300,000 Tweets every minutes with an average Tweet carrying information of the order of 200 bytes. These datasets have an associated time stamp and can be (loosely) located geographically; however alignment with existing data layer require customization on top of automating data ingestion to correctly localize the information. These unstructured datasets can enhance existing structured data layers (satellite, weather, etc) by offering contextual information regarding extreme events like flooding, wildfire, or other calamities much faster than any traditional data sources. Seamless integration of such information with established geospatial data layers (topography, road network, population density, etc.) requires the development of tools that can integrate data on the fly, index it [9,10], store it [4], and retrieve it on demand [12-13].

Using traditional database technologies to store geospatial data, have limited utility once the data volume exceeds a few Terabytes. Such data bases are also not well suited to handle geo-spatial data layers, as efficient indexing and joining, data layers have limited support.

Here we present a new geospatial big data platform, Physical Analytics Integrated Repository and Services (PAIRS), to process Petabytes of data and address the spatial and temporal complexity associated with heterogeneous data integration (Fig. 2). Historical and real time geospatial data sets are automatically downloaded, curated and stored in HBase table, which are then available for real time modeling

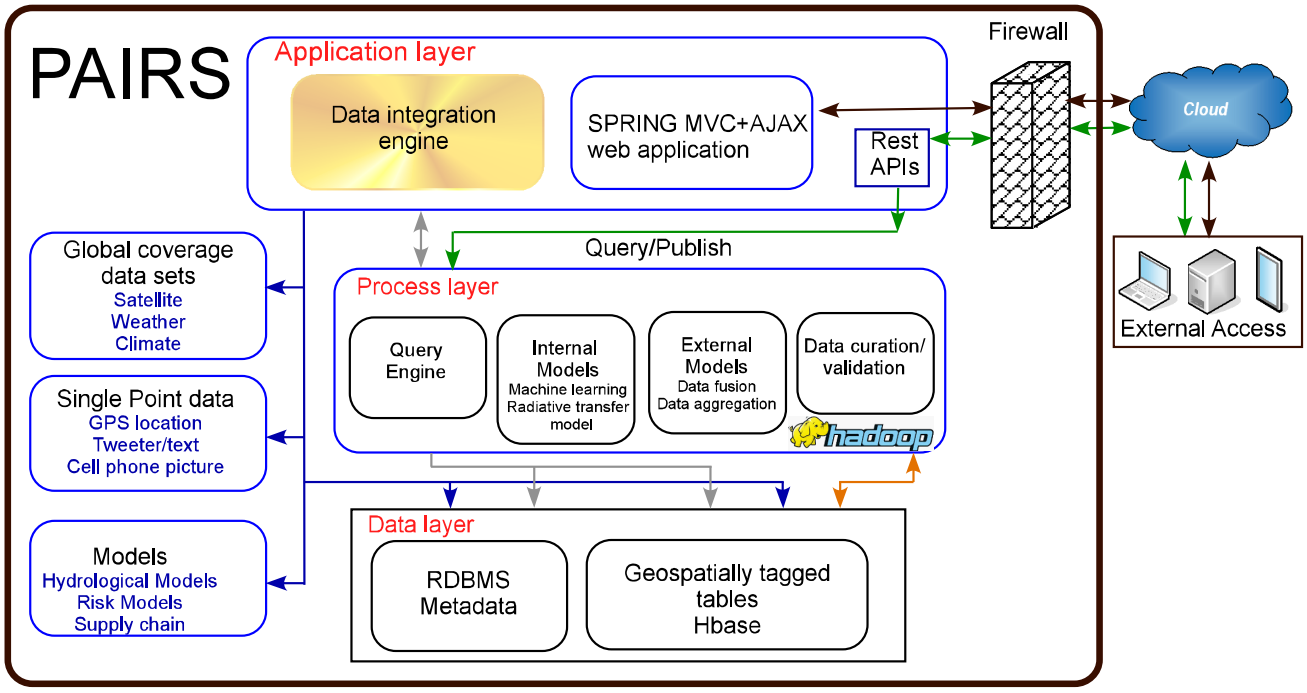

Figure 2. PAIRS architecture as a cloud service where a query retrieves (1) metadata from a traditional relational database (PostgreSQL) and (2) pulls geospatial data from HBase.

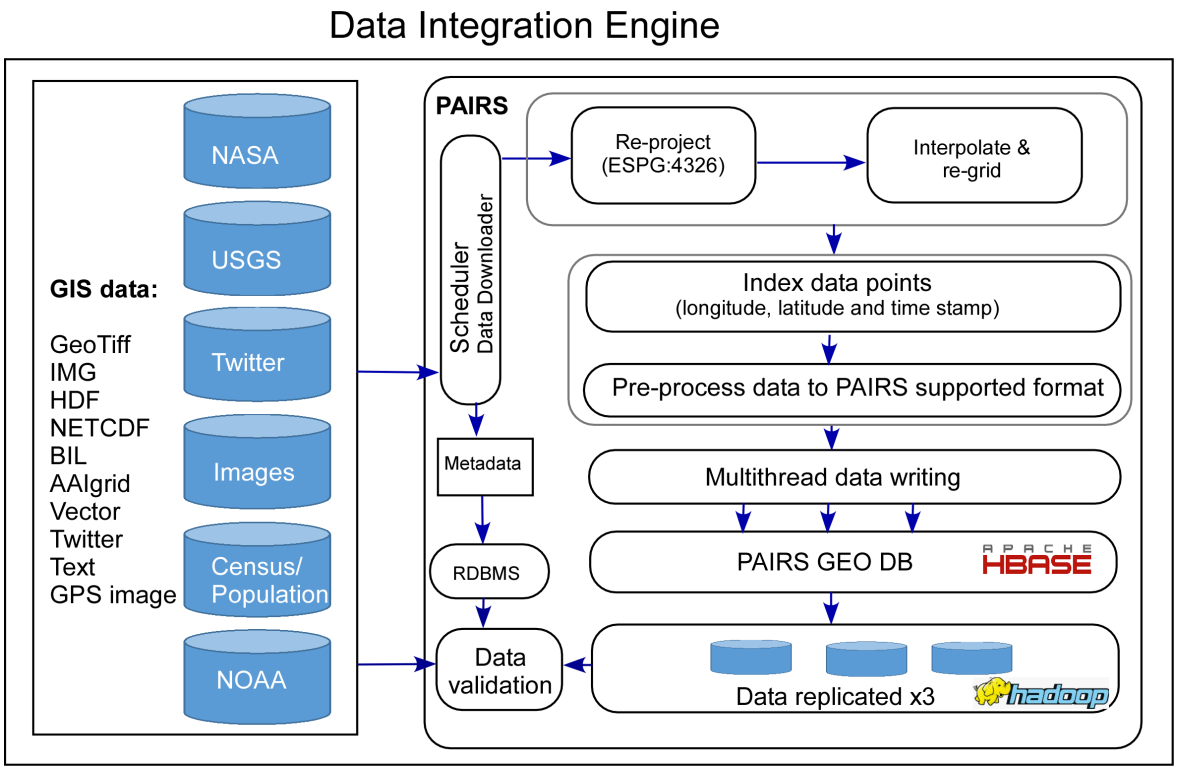

Figure 3. A detailed view of PAIRS's Data Integration Engine from Fig. 2: Geospatial data are automatically downloaded and processed. A multi-threaded process writes the data to HBase on top of Hadoop's HDFS. Data validation and verification for accuracy is automated to ensure that the best possible data layers are uploaded into PAIRS.

and analytics. Data curation encompasses; conversion to a common datum, and aligned on a well-defined spatial grid. The platform can be queried to retrieve data in multiple ways, like: (1) single point across large interval to create time series, (2) spatial query across an arbitrary sized area, and (3) filtered spatial and temporal query using a system of filters to retrieve subset of data from each layer.

PAIRS offers access to a repository of consistent historical and real time datasets that are aligned and indexed. It is developed on top of the open source big data technologies Hadoop and HBase [14]. It leverages MapReduce [15] to accelerate data queries by parallelizing search and data retrieval. One key differentiator of PAIRS is the multi-layer query capability, ability to search multiple data layers and filter then based on multiple search criteria where filters allows discovering locations or time periods that share the same characteristics in space and time. This capability provides a quick way to visualize in real time changes for a certain location and detect similarities or differences. In addition to existing data layers, any custom modeling or analytics layer can be uploaded into PAIRS.

II. Geospatial Data Management

Due to the complexity to describe earth surface and preserve accuracy on local scale, almost all geospatial data layer are provided in different projections, which makes data joining and alignment the most time consuming task of any geospatial effort. PAIRS integrates open source tools (GDAL, PROJ.4, etc.) to convert data layer projection to WGS84 coordinate system and facilitate map re-projections and data managements on the fly. These tools can operate on large datasets and they scale with data volume if they are parallelized. The WGS 84 (EPSG:4326) coordinate system is used for all data layers. Independent of the original formats, the platform converts all data layers to WGS 84, that is, it interpolates and re-grids the data before it is integrated into HBase storage (Fig. 3).

The main building block of PAIRS is the Data Integration Engine that handle data download, re-projection, and data indexing (Fig. 3). Any geospatial data format (raster, vector or geo-located images and text) can be integrated into the platform. If a data format is not raster, then it will be rasterized and handled as a large matrix. All data layers in PAIRS are aligned and snapped on a common grid, generated by continuously subdividing the global coordinate space [-180°, 180°] x [-90°, 90°] in terms of longitude and latitude (Fig. 4). The spatial resolution of PAIRS ranges from cm to hundreds of kilometers, with the smallest grid spacing of 0.000008˚ that corresponds to 0.8 m on the longitude. There are 26 grid layers with increasing lateral resolution (see Table 1).

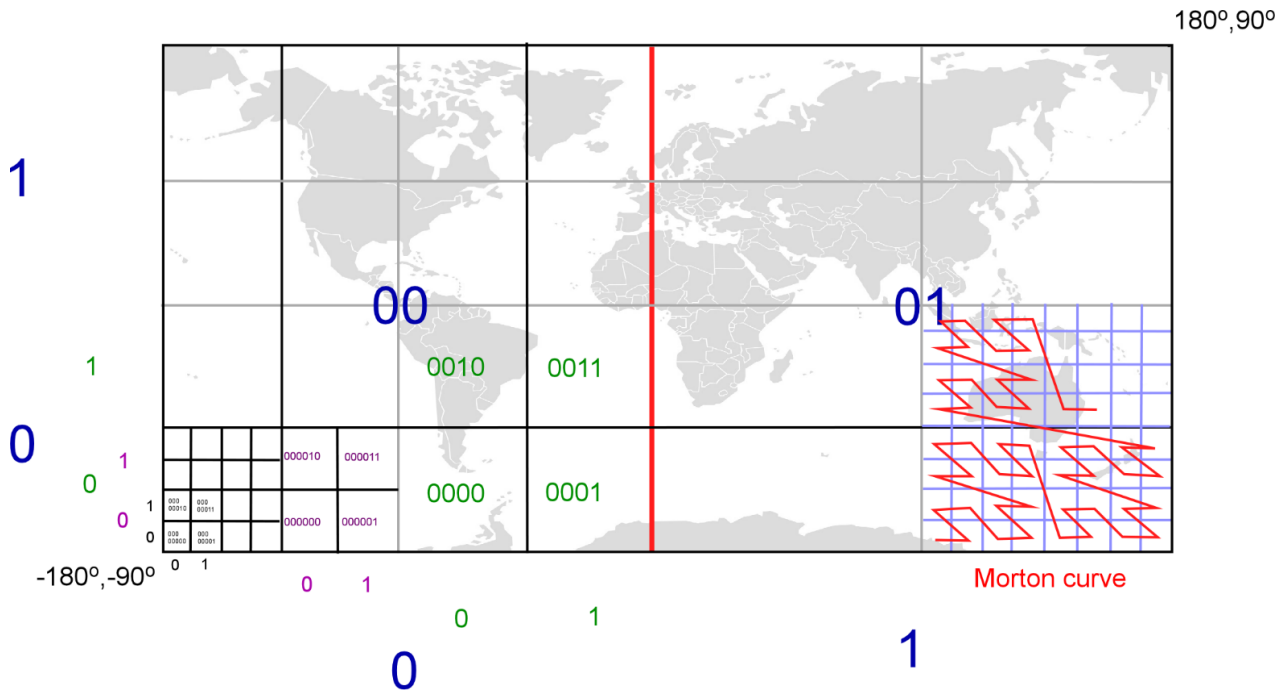

Figure 4. Global indexing of data in PAIRS: Recursive quad-tree scheme to partition two-dimensional data.

In order to adapt each datasets to PAIRS's resolution, the original data layers are interpolated, followed by rasterizing the data layer to the closest PAIRS grid cell size. The interpolation is chosen to preserve, as best as possible, data accuracy and distinctive features like shorelines, water bodies, road alignments, etc.

III. Data Indexing and Representation in HBase

Each data point for any data layer is indexed before is uploaded to the HBase table. The index includes a spatial location (longitude θ and latitude φ) and a timestamp t. Fig. 4 provides details about the indexing scheme implemented in PAIRS. The generation of the grid follows the well-known recursive procedure of a quad-tree, where the global map gets subdivided into 4 equally-sized cells labeled by the bits 00, 01, 10, 11. This process is repeated and the corresponding bits are appended to the right of their "mother" cell. The resulting bit sequence is interpreted as the binary representation of the z-index. Linking these indices with respect to increasing value on the two-dimensional map generates the Morton curve [9]. It targets at approximately preserving the "locality" property, i.e. nearby two-dimensional raster pixels stay close to each other when arranged on a line [10]. This property becomes important when designing HBase's row key which is stored in alphabetical order.

Using *m* as the number of recursions, the map can be divided into 4^m grid cells which implicitly sets the spatial resolution in degrees to $\Delta\varphi=90°/2^m$. In order to avoid rounding errors due to limited numerical precision, the resolution layer index n is introduced such that the resolution becomes $\Delta\varphi=2^{n+2}\cdot 10^{-6}$ (see Table 1). Note that the resolution in km is consistent along the longitude, but does change along the latitude, with larger values expected near the equator and smaller values near the poles.

The transformation from the pair (longitude, latitude) = (x,y) that matches the lower left corner of the grid cell to its z-index provides the first part of the row key k for the HBase table that stores all geo-spatial information. Since PAIRS stores datasets that have an associated time stamp, each index will have a temporal component defined as t in seconds, i.e. k = (z,t).

This scheme is similar to the approach of MD-HBase [16] which demonstrates a use case of geographically tracking moving objects in real time. In particular the focus is on efficiently scanning sparse data. However, the PAIRS system independently treats the time direction from the z-ordering - geographic raster layer pixels are static. Fox [17] presents another index structure to encode (geo-)spatial-temporal information into the key of the key-value store Accumulo [18], an Apache project.

Due to HBase's row key ordering [14] all timestamps t_0, t_1, t_2, ... of any geo-information data in PAIRS at fixed spatial position z_0 gets grouped together. HBase SCANs therefore quickly deliver time series of geo-information [14]. PAIRS

Resolution Layer (n)	Δθ, Δφ [degree]	Δy [km], Δx[km](ϕ=0°)	Δx [km](ϕ=40°)
1	0.000008	0.00089	0.00067
2	0.000016	0.00178	0.00134
3	0.000032	0.00356	0.00268
4	0.000064	0.00712	0.00536
5	0.000128	0.01424	0.01072
6	0.000256	0.02848	0.02144
…			
9	0.002048	0.22786	0.17152
…			
26	268.43546	29863.444	22481.469

Table 1. Global grid spatial resolution in degree for longitude (θ) and latitude (φ) and the corresponding resolution in km at the equator and at 40 degrees latitude, respectively.

performance is tuned to retrieve local information across large time interval. We note that in the conventional framework, time series retrieval from large geospatial data is constrained as large number of files need to be opened, scanned, localizing data points of interest for each data point.

Each data pixel referenced by *k* is stored in the corresponding HBase cell that is uniquely identified by *(k,c),* where *c* represents the information based on HBase's column family and column qualifier [5,6]. Aggregating geospatial pixels into a larger block to store them into a single HBase cell has two advantages: (1) balancing the query speed of fixed geo-spatial information for multiple timestamps and (2) balancing a geo-spatial region at a given time. In current implementation, 32 x 32 pixels are stored per HBase cell. To each block of geo-spatial pixels, a key is assigned, which is the same key of the contained pixels with the 10 least significant bits dropped. To each pixel, a unique key is assigned inside the 1024 array, calculated using only the 10 least significant bits, and stored into the HBase table as binary data. Each HBase table has multiple columns where each individual column stores a specific data parameter. For example, weather data like North American Mesoscale models (NAM), multiple columns are defined, where each column is assigned to a physical quantities like temperature, precipitation, ground pressure, etc.

Since MapReduce can execute jobs directly on HBase tables, analytics run on PAIRS can be parallelized. In addition, the simple fact that any HBase cell is stored if and only if the corresponding row key is written (HBase command PUT), PAIRS handles very efficiently sparse geo-spatial information.

IV. Data Processing

The current datasets available in PAIRS are listed in Table 2 and could be grouped in four large categories: satellite, weather, survey, and analytics. Each category contains several datasets:

1. satellite images: Landsat [5], MODIS [6]
2. weather products: Parameter-elevation Relationships on Independent Slopes Model (PRISM) [19], North American Mesoscale (NAM), California Irrigation Management Information System (CIMIS), Global Forecasting System (GFS)
3. survey data: elevation (NEM) [20], land use (CROPSCAPE) [21], Soil (gSSurgo)
4. layers that are part of the analytics services: blended weather forecast and evapo-transpiration / irrigation forecasting

Datasets		Origin	Frequency	PAIRS grid resolution (degree)
Satellite	Landsat 7	USGS	16 days	0.000256
	Landsat 8	USGS	16 days	0.000256
	Modis /Vegetation Product	NASA	16 days	0.002048
	Modis Surface Reflectance	NASA	2 days	0.002048
Weather	Global Forecasting System(GFS)	NOAA	Daily+60h forecast	0.524288
	North American Mesoscale Forecast System (NAM)	NOAA	Daily +10 day forecast	0.032768
	European Centre for Medium Range Weather Forecast (ECMWF)	EU	Daily +10 day forecast	0.0655
	PRISM data	U of Oregon	Daily	0.032768
	CIMIS data	State of California	Daily	0.016384
Survey	National Elevation Data	USGS	Static	0.000008
	Soil data (SSURGO)	USDA	Static	0.000008
	Cland use (Cropscape)	USDA	1 year	0.000256
Analytics	IBM Blended Forecast	IBM	Daily	0.032768
	IBM Evapo-transpiration/ Irrigation	IBM	Daily	0.032768

Table 2. Current data layers in PAIRS that are updated continuously and two real time analytics, weather forecasting based on machine learning and irrigation forecasting for crops.

Data curation encompasses calculation of value, adding data layers such as elevation derived product, like slope and aspect ratio, as well as correcting Normalized Difference Vegetation Index (NDVI) for atmospheric effects. In addition, each data layer is periodically verified for accuracy and consistency.

Due to the difference in data acquisition rate (cf. Fig. 3) as well as different repositories used for storing original data layers, PAIRS data agents are querying each repository to retrieve the specific data layers. Besides data download, PAIRS scans its HBase key-value store to identify missing timestamps. An automatic request is submitted to retrieve missing data layers.

A special case are data layers that have partial spatial coverage and are acquired at different moment in time, e.g. Landsat with a 16 days visitation time where images are acquired as 185 km by 180 km tiles along a swath of the Earth's surface. Data validation requires to verify if all the tiles at a given timestamp are present in the HBase table. The HBase table is scanned for a representative data point that defines the longitude and latitude of the center of that tile. In case a tile is missing, a data request is submitted to obtain the data. If the data is available on the host server, the data integration engine seamlessly connects the tile into the existing data fabric.

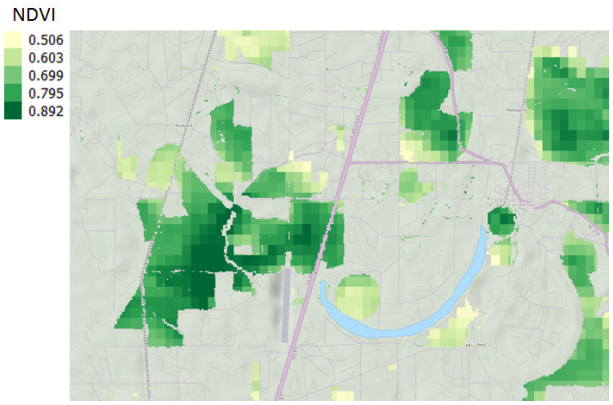

Figure 5. PAIRS multi-layer query that extracts farms in Mississippi growing corn and having NDVI larger than 0.5, where average weekly rain in September 2015 was larger than 10 mm.

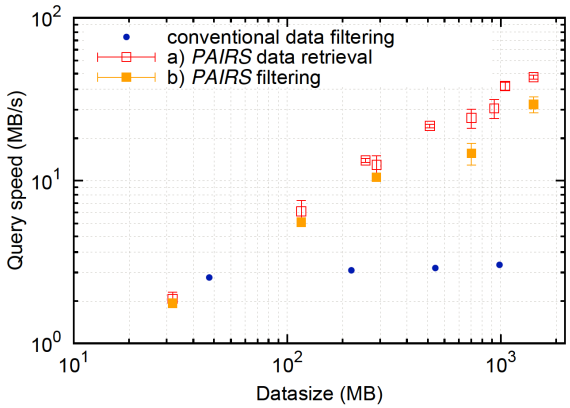

Figure 6. Performance metrics for conventional data filtering on a single machine using GDAL and Python vs. querying data from PAIRS: a) data retrieval and b) data retrieval including filtering.

V. Spatial-Temporal & Multi-Layers Queries

The simplest query that can be submitted to PAIRS is a time series query for a single data point defined by longitude and latitude where all existing data points acquired can be retrieved. A query for a single point, from a data layer, acquired daily for a 10 year period, can be retrieved in less than 1 second.

PAIRS supports range queries based on polygons for a given time interval. The PAIRS query API, implemented as a RESTful web service, automatically grabs the data acquired within the specified time interval and returns multiple georeferenced images tagged with the data acquisition timestamp.

A third type of query implemented in PAIRS is multi-layer queries where multiple layers can be queried and filtered based on specific parameters from one or multiple data layers. For example, one such complex query can be used to retrieve all farms that plant corn, whose mean precipitation across a week was larger than 10 mm, and where corn has NDVI values above 0.5. This query requires the platform to identify the geo-spatial locations where corn is planted, query the weather data layer and aggregate it for a week, filter all data layers based on the three conditions and return a data layer that fulfills all the above conditions. Fig. 5 illustrates an example of such a query. Note the visibility of the different resolutions of the Cropscape and NDVI layer that get joined for the result: While there exist isolated tiny spots of farms (cf. high resolution Cropscape layer), the interior of larger patches reveals the more coarse grained resolution of the NDVI layer.

For this query, the recursive procedure that generates the z-ordered index from the HBase row key k=(z,t) comes in handy. Scanning the 30 m resolution dataset for a specific value (e.g. Cropscape data layer for corn) one can directly employ the matching keys to investigate a layer with less resolution, e.g. precipitation data from PRISM. The reason being that if $z_a=<z_1z_2...z_p>$ is the p bits of the index labeling a given grid cell, cells of a finer resolution within this cell get assigned a corresponding z-ordered index according to $z_b=<z_1z_2...z_pz_{p+1}z_{p+2}...z_{p+q}.>$ with p+q bits. So, knowing z_b from corn data directly provides z_a for precipitation. PAIRS's indexing and aligned data layers accelerate multi-layer queries.

As pointed out in section III, PAIRS's indexing scheme builds on geo-spatial coordinates of the raster as well as temporal information. Therefore PAIRS exploits the HBase key design to perform range-queries over multiple timestamps. In a conventional scenario the user needs to deal with a collection of individual files representing a certain geospatial area over a certain time period. Fig. 6 shows (solid dots) the corresponding processing speed on a set of files using conventional tools on a single machine. The query time scales proportionally to the amount of processed data, i.e. the data retrieval speed is constant. In detail, the data is loaded from hard disk into memory. Then each pixel gets filtered by the condition: "value of the pixel is equal to a constant". For this task, the data need to be unzipped, re-projected and filtered – common steps that a geospatial data user will undertake if he processes the data.

PAIRS's querying speed (open & solid boxes) is faster in absolute value than conventional method, the benchmarking demonstrates increasing speed with increasing size of retrieved data - the hallmark of increased parallel processing due to the distribution of the data in the PAIRS cluster. Note, the major fraction of time in the conventional case is spent to unzip and re-project the data. PAIRS incorporates data compression and layer alignment by default.

Besides MD-HBase and an Accumulo based system there are multiple approaches to use open source software technologies in order to handle scalable solutions for big (geo)-spatial raster and/or vector data. To name a few, e.g. Hadoop-GIS [22] implements operations such as SQL-type JOINs on vector data (OpenStreetMap [23]) employing Apache Hadoop. Building on Hadoop as well, SpatialHadoop

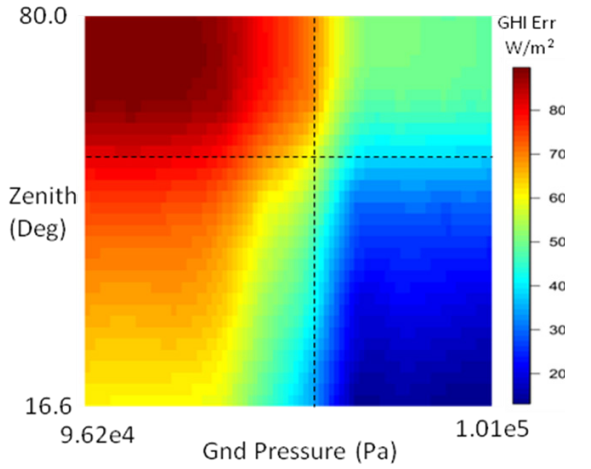

Figure 7. Global horizontal solar irradiance (GHI) forecast error dependence on forecasted ground level pressure and solar zenith angle.

[12] and its related implementations, SHAHED [24], TAREEG [25] & TAGHREED [25], query and visualize raster, vector and text data, respectively. An alternative geo-spatial big data platforms, ADAM [4] employs Apache Spark [32] to compete with traditional parallel image processing systems (cf. analytics on geo-spatial raster data) for astronomy. Yet another recent project, SpatialSpark [33] uses the Apache Spark technology to implement join operation on geo-spatial vector data. PAIRS particularly focuses on providing a whole, business-focused database and insight service to help to manage real-world geospatial data from multiple sources in a scalable manner on distributed compute resources.

VI. Analytics

A. Weather modeling/machine learning

PAIRS curates historical weather forecasts and thus provides new opportunities to enhance forecasting accuracy through advanced statistical learning techniques. Current weather forecasting models that either use statistical correction or combine forecasts from different numerical weather models (Dynamic Integrated Forecast systems [29] or the Model Output Statistics systems [30]) have the limitation that only a restricted amount of forecast data is preserved and those will have either a short span of few days or are not updated with new forecasts. In general the forecast data are not saved for further analysis or model refinement.

In contrast, PAIRS enables on-demand rapid retrieval of historical forecasts and associated measurements (from weather stations that provide real time data) going back historically for a few years. Such capability enables systematic characterization of forecast error of weather models and the dependence of the error on other forecasted weather parameters. This analysis can be carried out across

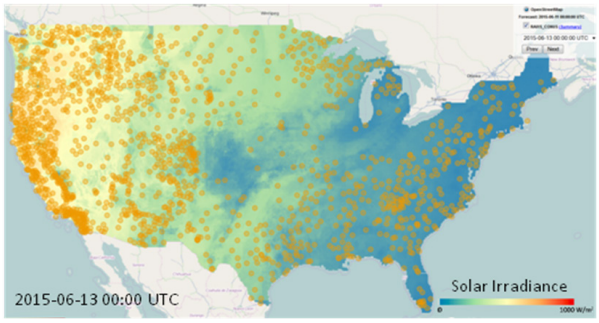

Figure 8. Global Horizontal Irradiance (color scale) forecast for 48 h in advance obtained via machine-learning based situation-dependent blending of the North American Mesoscale model and the Short Range Ensemble Forecast Model.

global sites to understand the performance of various weather models used for forecasting. As an illustrative example, Fig. 6 demonstrates one case of how global horizontal solar irradiance (GHI) forecast error is dependent upon its forecast of ground level pressure and solar zenith angle.

The GHI forecast is part of the of the widely-used North America Mesoscale (NAM) weather model that is frequently used in USA for business support including aviation safety etc. The result is obtained using Functional Analysis of Variance (FANOVA) for a Surface Radiation (Surfrad) Measurement station (BND station at Champaign, IL) averaged for period 2014-1-1 to 2014-12-28 as are discussed in [31].

Such analysis identifies the different "weather situations" (unique to specific locations) in which the forecast error of the NAM model is dependent on different physical parameters that are specific to geographical locations. For instance, the two parameters (ground pressure and zenith angle) apparently create four different weather situations pertaining to the GHI forecast error as indicated by the dashed lines in Fig. 7. In each of the dominant "weather situation," the forecast can be corrected based on individual weather model or combining different weather models using statistical learning methods to improve overall forecasting accuracy [13].

As an example, Fig. 8 shows a 48 hour ahead forecast of global horizontal solar irradiance (GHI) of contiguous US (CONUS) obtained via situation-dependent blending of forecasts from two weather models—the NAM model and the short range ensemble forecast (SREF) model. This forecast is issued at 2015-06-11 00:00 UTC for 2015-06-13 00:00 UTC. The model blending is trained by historical forecasts and measurements at ~1640 remote automatic weather stations (RAWs) of the MesoWest network (yellow circles).The historical forecasts and measurements of GHI and other meteorological parameter such as ground pressure and solar zenith angle for 60 days immediately before the forecast issuance time at ~1640 remote automatic weather stations (RAWs) (yellow circle in Fig. 8) are used as training data. For an arbitrary location in CONUS, its situation-dependent

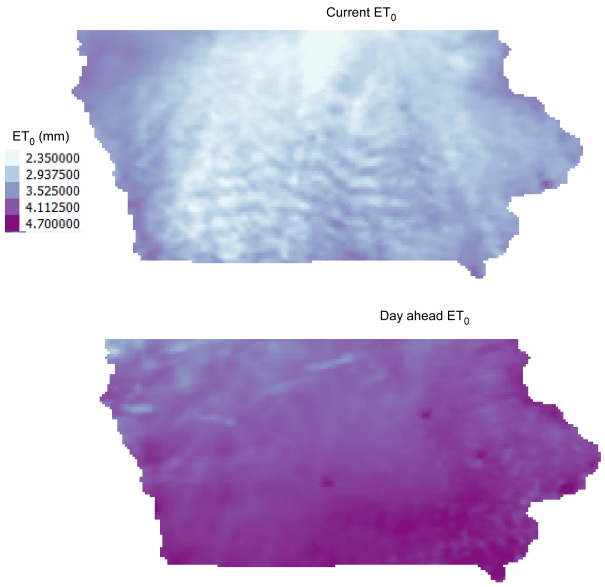

Figure 9. Spatial reference evapo-transpiration calculated from numerical weather model for current and day ahead on 09/01/2015.

model blending is determined using training data from ~10 to 15 neighboring RAWs sites. The method enables 30% improvement of GHI forecast accuracy, measured in terms of mean absolute error, compared to the two individual input weather models (NAM and SREF).

B. *Evapo-transpiration/Irrigation Forecasting*

Energy balance models can quantify the amount of water required by plants [32]. The evapo-transpiration is calculated as the difference between the input radiation from the sun and heat absorbed by soil and heat transfer from plant to air. These energy balance models can accurately calculate the right amount of irrigation that needs to be delivered to a plant in order to maintain the optimum amount of water in the crop [33]. The energy balance method developed at the spatial resolution of Landsat [34] combines (1) crop vegetation information from Landsat data and (2) satellite based soil temperature information, and (3) weather data integrated into reference evapo-transpiration using Penman-Monteith equation.

The energy balance method was demonstrated to give a very accurate assessment of water amount required for irrigation based only on Landsat data and weather data modeling [34]. One application of high resolution irrigation scheduling is in variable rate irrigation that can be implemented with central pivots or drip irrigation system. It was demonstrated that variable rate irrigation can save water while at the same time increase yield while maintain crop quality [35]. The advantage of satellite based evapo-transpiration calculation is that it can be scaled and translated to other geographies to provide consistent irrigation recommendations for all crop.

The irrigation forecasting is based on weather forecast for 10 days in advance. For evapo-transpiration, temperature, wind speed, solar radiation and relative humidity from NAM and GFS models are integrated with the Penman–Monteith [35] equation. In Fig. 9 the current and day ahead evapo-transpiration based on the NAM models is shown for the state of Mississippi calculated at a spatial resolution of 4 km.

The irrigation schedule can leverage additional data layers like precipitation that would enable to correct irrigation schedules by subtracting the from the irrigation recommendation. These corrections in irrigation schedule are very beneficial during extreme weather events or rainy seasons for better water management and optimum water allocations.

VII. CONCLUSIONS

Real time geospatial analytics requires large volume of geospatial data to be analyzed and processed. This requires a scalable platforms where data layers can be seamlessly incorporated into models. In this paper we presented the PAIRS platform, which is designed to accelerate analytics and provide access to curated datasets. PAIRS minimizes the "time to value" by providing curated data sets and eliminating data processing steps that normally should be run on data layers. Two analytics models are running on top of curated data sets: (1) improved and hyper localized weather forecast model based on machine learning, and (2) evapo-transpiration/irrigation forecasting. As more data layers are integrated in PAIRS, they will enrich existing data layers, and will enable more complex analytics to be run on top of the data.

ACKNOWLEDGMENT

We would like to acknowledge useful discussions with Stuart Siegel, Upendra Chitnis, Satish Gajjela, and Supratik Guha about geospatial data formats and analytics solutions.

REFERENCES

[1] M.C. Hansen et al., "High-resolution global maps of 21st-century forest cover change." Science 342, 850-853, 2013.

[2] E. Kalnay and C. Ming. "Impact of urbanization and land-use change on climate." Nature 423, 528-531, 2003.

[3] Q.Weng, "A remote sensing? GIS evaluation of urban expansion and its impact on surface temperature in the Zhujiang Delta, China." International journal of remote sensing 22,10, 1999-2014, 2001.

[4] F.A. Nothaft, M. Massie, T. Danford, et al, "Rethinking Data-Intensive Science Using Scalable

Analytics Systems," Proc. ACM SIGMOD, Int. Conf. Management of Data, 631-646, 2015.
[5] D.P. Roy et al. "Landsat-8: Science and product vision for terrestrial global change research." Remote Sensing of Environment 145: 154-172, 2014.
[6] MODIS, 2015, https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod13q1
[7] D. Starosta, "Data Processing, Product Generation and Distribution at the NWS National Centers for Environmental Prediction," 2012, http://www.nist.gov/itl/ssd/is/upload/Big-Data_ncep.pdf (retrieved on 09/18/2015).
[8] B. Raoult, "Big Data at ECMWF: Providing access to multi-petabyte datasets," http://www.copernicus.eu/sites/default/files/library/Big_Data_at_ECMWF_01.pdf (retrieved on 09/18/2015).
[9] G. Morton, "A computer-oriented geodetic data base and a new technique for file sequencing," IBM Canada: Unpublished report, 1966.
[10] M.F. Mokbel, and W.G. Aref, "Analysis of Multi-Dimensional Space-Filling Curves," GeoInformatica, vol. 7, 179-209, 2012.
[11] K.E. Taylor, R.J. Stouffer, and G.A. Meehl, "An overview of CMIP5 and the experiment design," Bull. Amer. Meteor. Soc., vol. 93, 485–498, 2012.
[12] A. Eldawy, and M.F. Mokbel, "SpatialHadoop: A MapReduce framework for spatial data," Proc. IEEE Int. Conf. on Data Engineering, 2015.
[13] R.T. Whitman, M.B. Park, S.A. Ambrose, and E.G. Hoel, "Spatial Indexing and Analytics on Hadoop," in SIGSPATIAL, 2014.
[14] L. George, "HBase: The Definitive Guide," O'Reilly Media, 1st ed., 2011.
[15] J. Dean, and S. Ghemawat "MapReduce: Simplified Data Processing on Large Clusters," OSDI '04, 2004.
[16] S. Nishimura, S. Das, D. Agarawal, A.E. Abbadi, "MD-HBase: design and implementation of an elastic data infrastructure for cloud-scale location services", Distrib Parallel Databases, vol. 31, 2013, pp. 289-319.
[17] A. Fox, Ch. Eichelberger, J. Hughes, S. Lyon, "Spatio-temporal Indexing in Non-relational Distributed Databases", IEEE Int. Conf. Big Data, 2013.
[18] https://accumulo.apache.org/
[19] C. Daly, R.P. Neilson, and D.L. Phillips," A statistical-topographic model for mapping climatological precipitation over mountainous terrain", J. Appl. Meteor., 33, 140-158, 1994.
[20] http://ned.usgs.gov/
[21] W. Han, Z. Yang, L. Di, P. Yue, "A geospatial Web service approach for creating on-demand Cropland Data Layer thematic maps", Transactions of the ASABE, 57(1), 239-247, 2014.
[22] A. Aji, et al. "Hadoop GIS: a high performance spatial data warehousing system over mapreduce." Proceedings of the VLDB Endowment 6,11,1009-1020, 2013.
[23] http://www.openstreetmap.org
[24] A. Eldawy, M. F. Mokbel, S. Alharthi, A. Alzaidy, K. Tarek, and S. Ghani, "SHAHED: A MapReduce-based System for Querying and Visualizing Spatio-temporal Satellite Data," in ICDE, 2015.
[25] L. Alarabi, A. Eldawy, R. Alghamdi, and M. F. Mokbel, "TAREEG: A MapReduce-Based System for Extracting Spatial Data from Open- StreetMap," in SIGSPATIAL, Dallas, TX, 2014.
[26] A. Magdy, L. Alarabi, S. Al-Harthi, M. Musleh, T. Ghanem, S. Ghani, and M. F. Mokbel, "Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs," in SIGSPATIAL, 2014.
[27] H. Karau, A. Konwinski, P. Wendell, M. Zaharia, "Learning Spark", O'Reilly Media,1st ed., 2015.
[28] http://simin.me/projects/spatialspark/ (retrieved on 09/24/3015)
[29] W. Myers, G. Wiener, S. Linden, S.E. Haupt, "A Consensus Forecasting Approach for Improved Turbine Hub Height Wind Speed Predictions," American Wind Energy Association (AWEA), 2011.
[30] R. Glahn, K. Gilbert, R. Cosgrove, D.P. Ruth, and K. Sheets, "The Gridding of MOS," Weather Forecasting, vol. 24, 520–529, 2009.
[31] S. Lu, Y. Hwang, I. Khabibrakhmanov, et al, "Machine Learning Based Multi-Physical-Model Blending for Enhancing Renewable Energy Forecast – Improvement via Situation Dependent Error Correction," In press, Proc. Euro. Control Conference, 2015.
[32] R.G. Allen et al., "Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56," FAO, 300, D05109, 1998.
[33] W.G.M. Bastiaanssen et al. "A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation," Journal of hydrology, vol. 212, 198-212, 1998.
[34] R.G. Allen, M. Tasumi, R. Trezza, Satellite-based energy balance for mapping evapotranspiration with internalized calibration (METRIC)—Model, Journal of irrigation and drainage engineering 133,4, 380-394, 2007.
[35] L. Sanchez, M. Mendez-Costabel, B. Sams, A. Morgan, N. Dokoozlian, L. J. Klein, N. Hinds, H. F. Hamann, A. Claassen, D. Lew " Effect of a Variable Rate Irrigation Strategy on the Variability of Crop Production in Wine Grapes in California." ISPA Conference proceeding June, 2014.