netherlands eScience center by SURF & NWO

Converging Sciences Call
Grant application form 2012

REGISTRATION FORM (BASIC DATA)

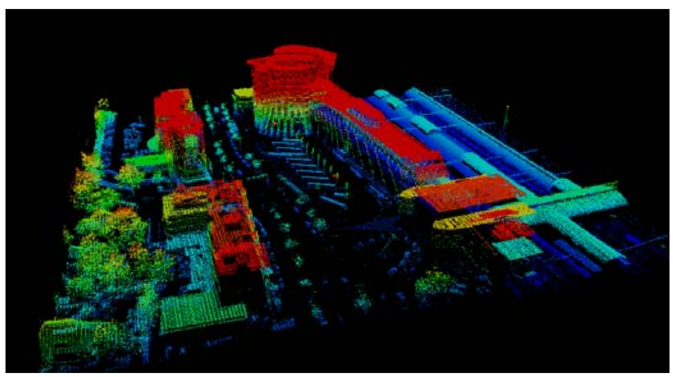
1 Details of the applicant(s)

Principal Investigator

Name, first name, title(s)	van Oosterom, Peter, prof.dr.ir.		male
Date of birth	30 July 1963	Date of PhD	20 December 1990
Position	professor		
End contract	Fixed/permanent position (Vaste aanstelling geen einde contract)		
Affiliation	TU Delft		
Department	OTB Research Institute for the Built Environment	Section	GIS technology
Postal Address	Jaffalaan 9	Zip/city	2628 BX Delft
Tel / Fax	015-2786950 / 015-2784422	E-mail	P.J.M.vanOosterom@tudelft.nl

Co-applicant (copy and paste if needed)

Name, first name, title(s)	van Wezenbeek, Wilma, ir.		female
Date of birth	7 June 1967	Date of PhD	
Position	other: director		
End contract	Fixed/permanent position (Vaste aanstelling geen einde contract)		
Affiliation	TU Delft		
Department	Library	Section	
Address	Prometheusplein 1	Zip/city	2628 ZC Delft
Tel / Fax	015-2787638	E-mail	W.J.S.M.vanWezenbeek@tudelft.nl



2a. Title of the proposal
Massive point clouds for eSciences

2b. Keywords
Data management,
point clouds,
geo-information,
levels of detail,
streaming data

3 Classification
You are requested than one theme, please indicate all the themes your research belongs to. If your research fits no theme, but it fits the aim of the call, then please shortly explain the connection with the aim.

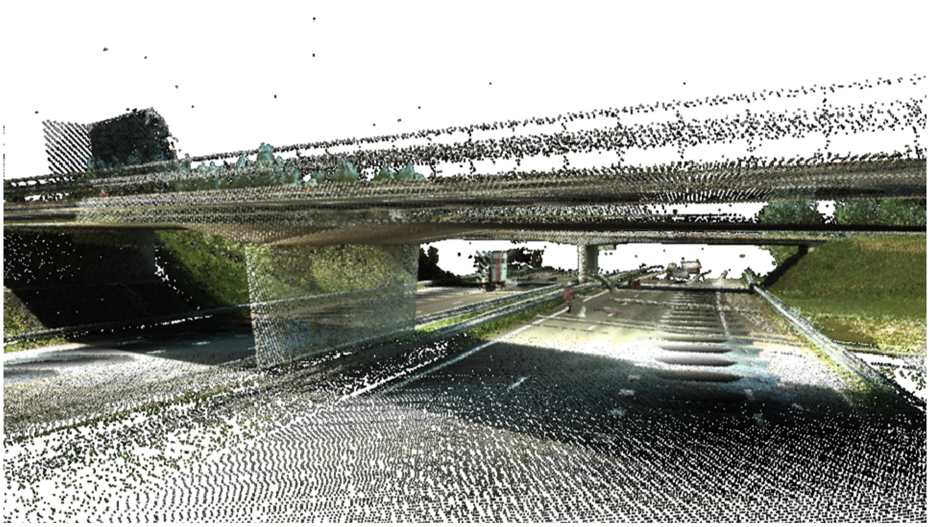
☐ Chemistry & Materials
☐ Humanities & Social Sciences
☐ Life Sciences
☒ Sustainability & Environment
☒ eScience Methodology & “Big Data”
☐ Other theme, please explain shortly

4. Composition of the Research Team

Name	Initials	Titles	Role in the project	Funded by
Van Oosterom	P.J.M.	Prof.dr.ir.	TU Delft, GIS technology, project leader	project
Tijssen	T.P.M.	Drs.	TU Delft, GIS technology, geo-database expert, researcher	project
Suijker	P.M.	Drs.	TU Delft, Library, contact with research & education users, dissemination & disclosure of point cloud data	project
Rombouts	J.P.	Ir.	3TU.Datacentrum, Long-term provision of ICT-infra (hardware/expertise)	project
Nouwens	H.	Ir.	TU Delft Shared Service Center ICT, storage facilities	project
Ravada	S.	Dr.	Oracle spatial, New England Development Centre (USA), improving existing software	Oracle
Alkemade	I.	Ir.	Rijkswaterstaat, data owner (and in-house applications)	Rijkswaterstaat
Kodde	M.	Ir.	Fugro, point cloud data producer	Fugro
eScience Engineer			eScience Engineer from NL eScience Center, designing and building ICT infrastructure	project

The three non TU-Delft partners in this project (i.e. Rijkswaterstaat, Fugro and Oracle) will co-finance this project by delivering their services and expertise, and by providing access to data and software (development). Their strong motivation for and their important contributions to this project are stated in their commitment letters accessible via the active links below:

1. Commitment Letter Rijkswaterstaat: http://www.gdmc.nl/oosterom/eSciencePC_RWS.pdf
2. Commitment Letter Fugro: http://www.gdmc.nl/oosterom/eSciencePC_Fugro.pdf
3. Commitment Letter Oracle: http://www.gdmc.nl/oosterom/eSciencePC_Oracle.pdf



netherlands
eScience center
by SURF & NWO

Converging Sciences Call
Grant application form 2012

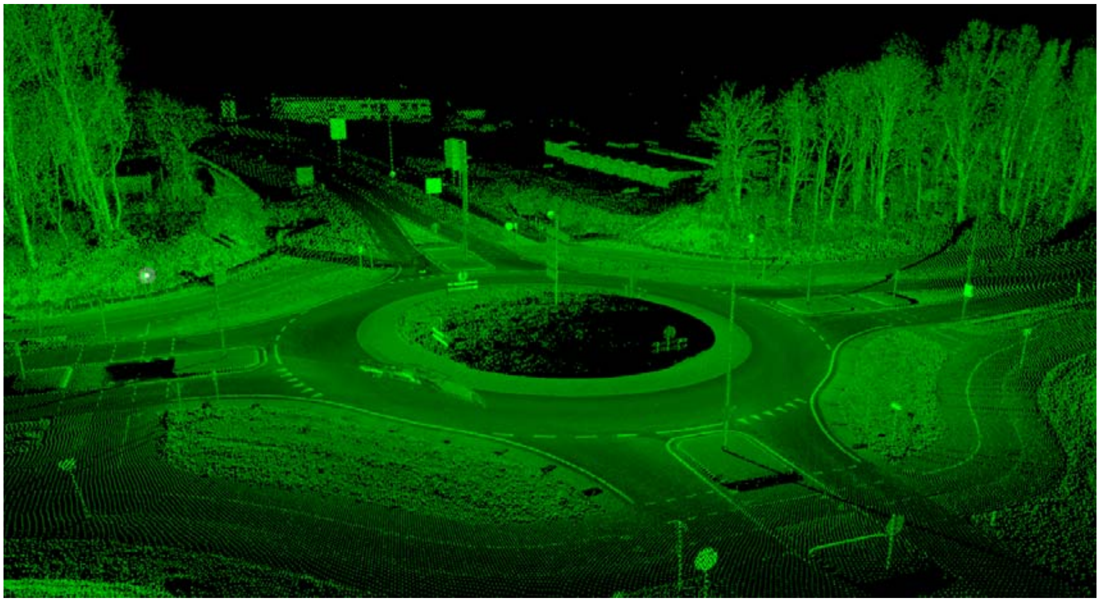
5a Summary of research proposal

Modern geographic data acquisition technologies generate point clouds with billions (or even trillions) of elevation/depth points. One example is the height map of the Netherlands (the AHN2 dataset). The point clouds have attracted a lot of attention from other research disciplines, such as: flood modeling, dike monitoring, forest mapping, generation of 3D city models, etc. However, the main problem with these point clouds is that they are simply too big (several terabytes) to be handled efficiently by common ICT infrastructures. Due to the lack of (data management, dissemination, processing and visualization) tools, the data are not being used to their full potential. Within this project several novel and innovative eScience techniques will be developed: [1] parallel ICT architecture, [2] new core support for point cloud data types in the spatial DBMS, [3] Web Point Cloud Service protocol (WPCS, progressive transfer from server to client based on multi-resolution representation), [4] coherent point cloud blocks based exploiting spatial clustering & indexing, [5] Point cloud compression (storage & transfer), [6] caching strategy, [7] exploit the GPU at client side and [8] fine tuning of all mentions parts within the complete system. Our work will also result in proposals for new standards: a. SQL/SFS extension for point clouds and b. web services (WPCS) to OGC and/or ISO TC211. The goal is a scale-able (much more data and users without architecture change) and generic solution (keep all current standard object-relational DBMS and integrate with existing spatial vector and raster data functionality).

5b. Summary for the general public (preferably in Dutch)

Nederlandse titel van het voorstel: Extreem grote puntenwolken voor eSciences

Hedendaagse geografische data-acquisitie technologieën genereren enorme puntenwolken bestaande uit miljarden (of zelfs biljoenen) hoogte & diepte punten. Een voorbeeld is de hoogtekaart van Nederland: de AHN2 dataset. Deze dataset trekt veel aandacht van verschillende onderzoekdisciplines, om bijvoorbeeld overstromingen te modelleren, dijkbewaking te verscherpen, bos in kaart te brengen of 3D-stadsmodellen te genereren. Het grootste probleem met puntenwolken is echter dat ze te groot zijn om efficiënt te worden gebruikt binnen de gangbare ICT-infrastructuur. Dit project ontwikkelt verschillende nieuwe 'big data' ICT- technieken om toegang en gebruik van puntenwolken te verbeteren en hiermee eScience op genoemde disciplines te ondersteunen.



RESEARCH PROPOSAL

6 Description of the proposed research

6a Science including relevance to eScience

More and more scientists need a scientific geodata infrastructure to obtain further and effective research progress (Craglia, 2012). This is true for the traditional ‘geo-information related sciences’ (geology, spatial planning, geography, geodesy, cartography, geophysics), but also for a growing number of other ‘non geo-information sciences’ such as criminology, medicine (cause of diseases and spread of viruses), natural language spread and development, water management, history/ archeology, etc. One of the biggest challenges is the efficient handling of the huge amounts of data, and point cloud data form a prime example of this. The importance of these data was also recently demonstrated, via a visualization of a small point cloud data subset to illustrate flood simulation, to Neelie Kroes (Vice-President of the European Commission responsible for the Digital Agenda) in the morning of 3 sept’12 before the opening of the academic year at the TU Delft. In (Kroes, 2012) she indicated that ICT solutions for handling ‘big data’, complex experiments and collaboration are crucial for the future of science.

Modern data acquisition technology for topographic information increasingly relies on so called point clouds. These point clouds are collected by applying new technologies such as laser scanning from airborne, mobile or static platforms, dense image matching from photos or multibeam echo-sounding. These modern data acquisition systems permit us to collect billions (trillions) of elevation/depth points for a given area, very quickly and with unprecedented precision (van Oosterom et al., 2010).

One example of the use of this technology in the Netherlands is the effort to build an accurate height map of the whole country containing as much as 6 to 10 samples per square meter (the AHN2 dataset; see illustration on page 1). Recently, such datasets have attracted a lot of attention because they are of great importance for several key applications related to the environment, such as: flood modeling, dike monitoring, forest mapping, generation of 3D city models, etc. It is expected that AHN3 will feature an even higher point density (as already in use at some today; e.g. Rotterdam), because data acquisition systems have evolved significantly since the start of the AHN2 project.

Within the TU Delft the university library is the distribution point for (geo-) data, including AHN2. The current users of AHN2 include a range of Faculties: Architecture (Urbanism, Landscape architecture, Ing. Steffan Nijhuis), Civil Engineering and Geosciences (e.g. Dr. ir. Olivier Hoes of Water Management and Dr.ir. Dominique Ngan-Tillard of Geo-engineering), Aerospace Engineering (Mathematical Geodesy & Positioning, prof.dr.ir. Ramon Hanssen), and Electrical Engineering, Mathematics and Computer Science (Computer Graphics & Visualisation, Dr.ir. Gerwin de Haan). ‘Outside’ the TU Delft (but on campus), the library also supports users at Deltares. Also more and more students are using this data.

However, the main problem with these datasets, as well for the Library as the academic users, is that they are simply too big to be handled efficiently with common ICT infrastructures. With the software tools currently available to practitioners, it is very difficult to simply visualize complete data sets, and the manipulation and processing of the data is nearly impossible. The situation can be summarized as follows: as far as point cloud datasets are concerned, current technology enables creation of data at a much higher rate than regular (or average) research infrastructure can process them. This obviously hinders the use of the data by practitioners (note that this is true for both academic users and non-academic users, e.g. within Rijkswaterstaat).

Considering all the point cloud datasets collected and stored for the area of the Netherlands (both for bathymetric and land-use purposes), there are several terabytes of extremely precise information available. This information could be beneficial to countless applications, but unfortunately, because of a lack of (processing) tools, most of these datasets are not being used to their full potential (for example, by first reducing the data into a 0.5m grid or 5m grid, the data is losing potentially significant detail). Basically, we are sitting on a gold mine but we are not exploiting it. As will be described below, several novel eScience techniques will be studied and developed to improve the access to and ease of processing of this data.

6b Approach and methods

The main objective of this innovative initiative is to develop an infrastructure for the storage, the management, the analysis and processing, the dissemination, the visualisation and the manipulation of massive point clouds. This is done by [1] creating parallel ICT architecture solutions with well-balanced configurations (including hardware, operating system and database management software), and [2] new core support for point cloud data types in the existing spatial DBMS -besides existing vector and raster data types-, capable of exploiting the parallel system architecture. The solution further includes: [3] a specific web-services point cloud protocol (supporting streaming, progressive transfer based on Level of Detail (LoD)/multi-resolution representation in form of a data pyramid); [4] coherent point cloud blocks/pages based on spatial clustering & indexing of the blocks and potentially within the blocks; [5] point cloud compression techniques (storage & transfer; e.g. use compact 'local block' coordinates and compact binary encoding); [6] a point cloud caching strategy at client (supporting panning and zooming of selected data); [7] exploit the GPU's (at client side) given the spatial nature of most GIS related computations [8] and fine tuning of all mentioned parts of the solution within the overall system. This innovative infrastructure will be able to handle these very large datasets and enable a large community to work with it by reducing bandwidth and processing power requirements for clients to a minimum.
The AHN2 dataset of the Netherlands will be used as a serious test case. The multi-resolution/LoD influences many of the aspects: storage, dissemination, visualization, analysis, etc. For data dissemination the current practice of sending external hard-disks by mail will be replaced by a Web Point Cloud Service (WPCS – as described in WP2 under section 7). This in a manner similar to the other operational services based on ISO's Web Map Service (WMS for raster geo-data) and Web Feature Service (WFS for vector geo-data). It will be investigated how such a Web Point Cloud Service can work in conjunction with the defacto standard for compressing laser data LASzip (Isenburg, 2011) and the new ASTM standard E57 (ASTM, 2011).
The platform built will contribute to and increase the research topics related to the (AHN2) massive point cloud by also providing more 'higher level' point cloud functions; e.g. derivation of slope/aspect ('surface' normal vectors), conversion to a grid format, detection and removal of double points, calculations of area/volumes, creation of simplified height map, extraction of basins, nearest neighbour analysis, etc. Within the project scope is also the development of client-side (end user) software, including interactive selection techniques and data visualisation algorithms.
Work package 1 (WP1) will focus on the requirements from all different users (see section 7). Big Data requirements will have a great demand on the reliability, availability, scalability and performance (RASP) of the underlying infrastructure, due to the sheer size or complexity of the data. Understanding these requirements while pertaining agility, flexibility, openness and responsiveness and being able to utilize modern cheap and fast hardware, storage and memory is quite a challenge. This requires an application platform and computing foundation, which is capable of handling both specific and generic tasks.
For the parallel hardware and DBMS it is necessary to select a middleware that can be easily adjusted to the specific requirements of point clouds and other (spatial) data, making it possible to adapt (or even develop) the different components independently if necessary. The selection of the middleware is influenced by many aspects, such as the specific necessities of the DBMS, the hardware infrastructure and the transfer of data between the DBMS and the parallel hardware. The result will be an architecture description that has the following characteristics:

- a basic platform consisting of two layers
    - *Virtualization* is a technique for hiding the physical characteristics of computing resources from the way in which other systems, applications, or end users interact with those resources.
    - *Containers* provide a runtime platform and provide common services like transaction support, management, security, and network communications that can be leveraged by the applications.
- a computing foundation consisting of the 3 layers:
    - *Distributed computing* allows multiple, autonomous computers to work in concert to solve a problem or provide a business solution.
    - *Grid computing* provides the ability to pool and share physical resources. It is a form of distributed computing that allows software to run on multiple physical machines in order to achieve availability and scalability requirements.
    - *Caching* increases performance by keeping data in memory rather than requiring the data to be retrieved from external storage.

From these characteristics a *Big Data reference architecture* is proposed as an overarching framework for the 'massive point cloud for eSciences' project. This reference architecture is generic enough to meet other Big Data project requirements (but will be made more specific to meet the project specific requirements; see section 8 Instrumentation) and includes:

i. High Level Architecture
ii. Service Exposure
iii. Service Flexibility
iv. Durability Requirements

i. High Level Architecture

The High Level Architecture consists of three layers (see Figure 1). Layer I provides the end user view on the data, a typical graphical client. Layer II provides the services on three sub-layers:

a) II-1 provides end user services, commonly known as business services. Protocols like JSON, SOAP and REST
b) II-2 provides composite services, these combine II-3 data or objects and have knowledge about GIS functions.
c) II-3 provides simple access to the basis data or objects retrieved from Layer III.

Layer III provides High Volume, Selectivity, Clustering, Standard Database Technology & Appliances. During the project this is implemented by the suite of Oracle products (see section 8 Instrumentation). The layer interoperability will be Open Standards Based.

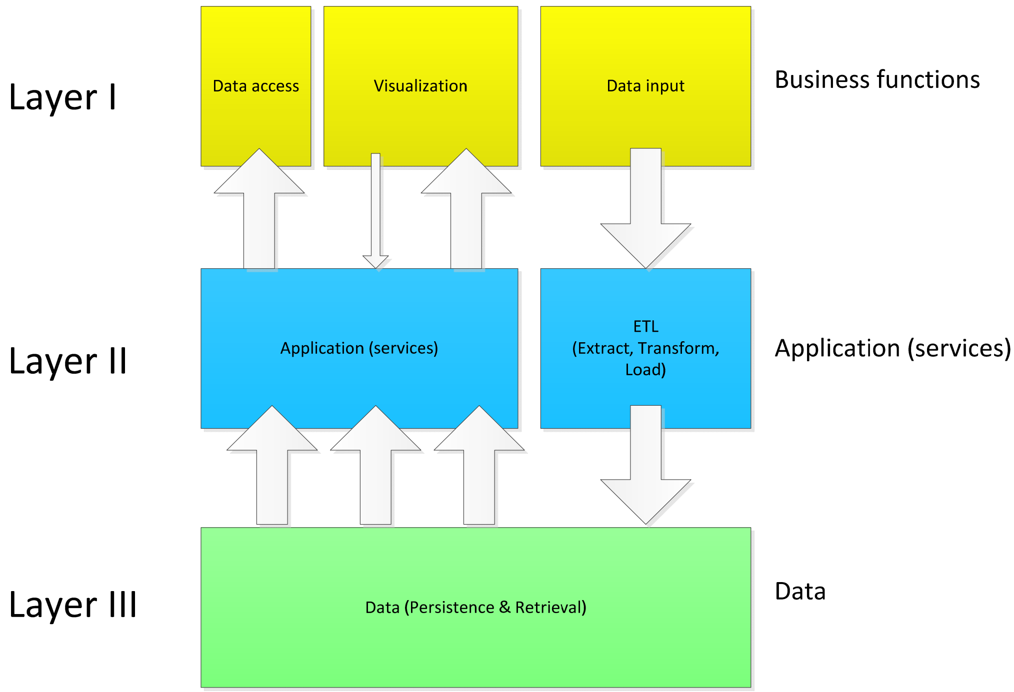


Figure 1. high level architecture

ii. Service Exposure

The services from the sub-layers from layer II are all exposed (see Figure 2). This allows for various types of reuse of the services.

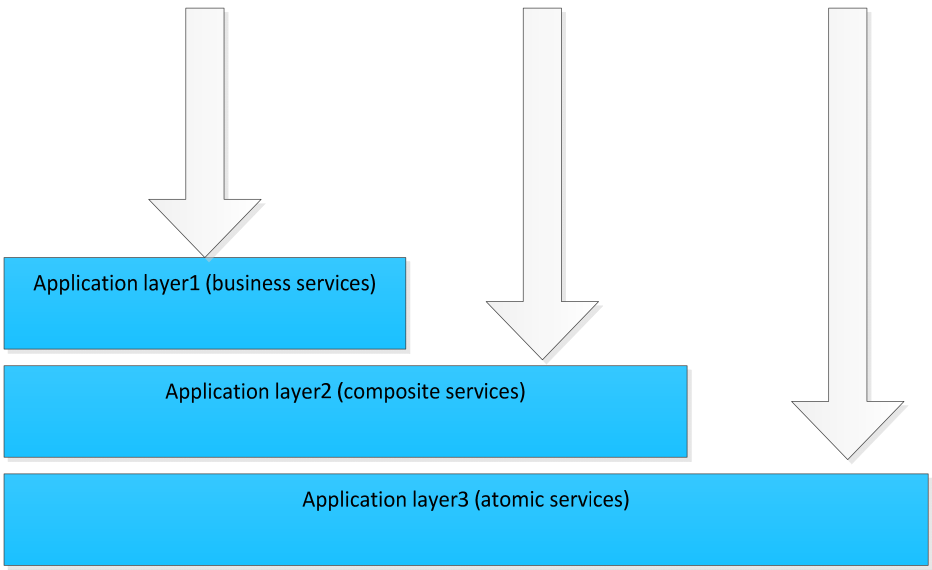


Figure 2. sub-layer services exposure.

iii. Service Flexibility

Application layer II can be implemented in various ways.

- A cluster of web services (small tasks, quick response)
- A HADOOP cluster with shared persistence, importing and manipulating large volumes of data (big jobs, one time only)
- Specialized ETL tools (big jobs, one time only)

The needed capacity will be shifted horizontally or possibly scaled out to external resources such as cloud instances (see figure 3).

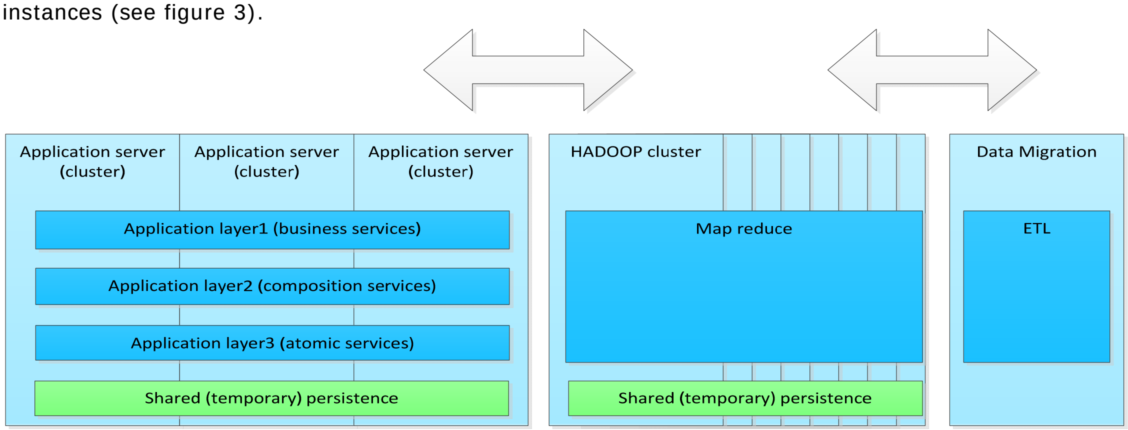


Figure 3. Service Flexibility

iv. Durability Requirements

The different layers have different requirements for their durability (with focus on a hard and software stack that is financially sustainable and technically maintainable). On the one hand this allows for replacement of the components used in the layers, on the other hand demands a stable choice.

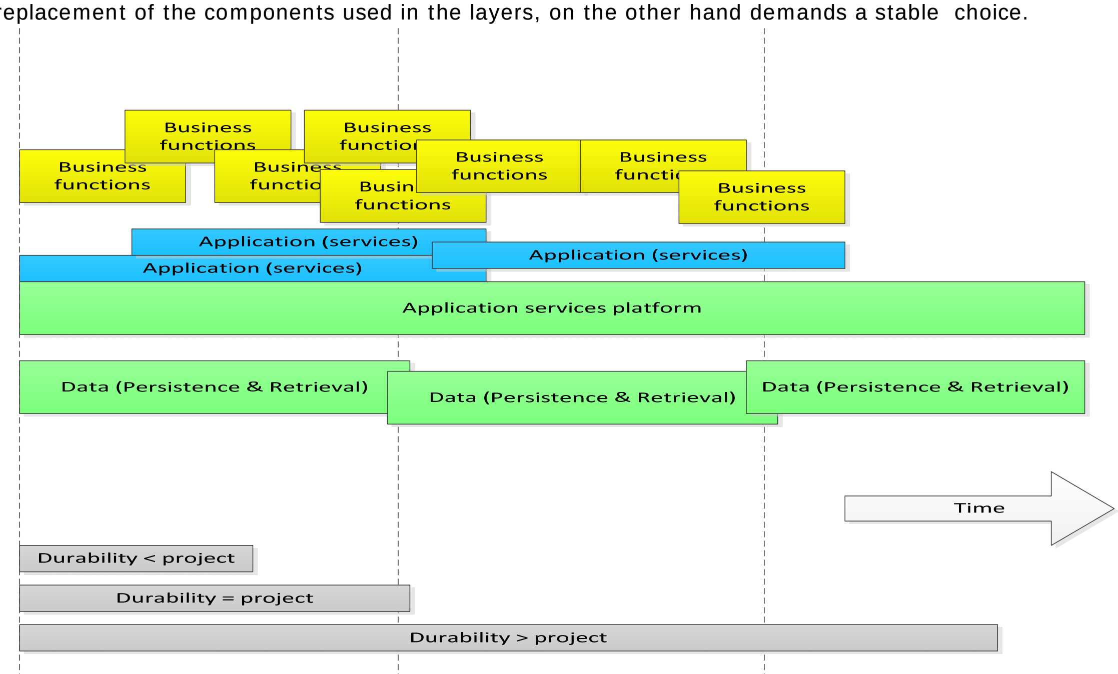


Figure 4. durability requirements

The implementation of this architecture will result in a scale-able and generic solution:

- Scale-able because the solution works well with 100 times more data and 100 times more users without architecture change.
- Generic because it keeps all current standard object relational Database Management Systems and spatial functionality; it integrates existing raster and vector data support with the new point cloud functions and combines it with advanced data transfer technologies; e.g. the selection, transfer and visualization of points that are within 1 meter distance buildings (stored in database polygons) in Delft and more than 150 years old.

6c Sustainability, dissemination and collaborations

The results of this project will be integrated in and hosted by the 'data dissemination' environment of the TU Delft Library: the TU Delft map room (http://www.library.tudelft.nl/en/visitor-info/facilities/map-room/) and the 3TU.Datacentrum (http://datacentrum.3tu.nl/en/about-3tudatacentrum/). The project results should serve the academic users of large point cloud data sets (e.g. AHN2), much better than today. The eScience engineer of NLeSC will cooperate with the GIS technology experts from the TU Delft in cutting edge R&D (supported by Rijkswaterstaat, Fugro and Oracle) resulting into the above mentioned desired services that will make new research possible.

The intention is not to limit the use of these services to the TU Delft users, but open these services to all academic users in the Netherlands (in line with the goals of UKB werkgroep 'Kaarten & GIS' and if successful also the Maps4Science initiative; see http://www.maps4science.nl).

The project team also cooperates with the software industry, i.e. Oracle. Oracle database spatial extension is currently the data management platform offering the most spatial functionality. Oracle (spatial) is also the current solution for managing (other) types of spatial data at the TU Delft library (and other Universities in the Netherlands). Oracle participates with the intention that the initial, but currently too limited performance of the Oracle point cloud solution, SDO_PC (Wijga-Hoefsloot, 2012) will be improved by integrating the results from this project.

Besides the cooperation with project partners, more initiatives are taken to form a national consortium of Geodata users and specialists. The willingness to cooperate is very high which will support the use and further development of the services developed within this project; e.g. see the 10 Maps4Science partners (http://maps4science.nl/partners/) and the 39 supporters (http://maps4science.nl/supporters/) including their support letters.

In WP8 'Investigate operational model for facility after the project' (see section 7) three possible lines of sustainability will be further explored:

1. Within research (TUD Library, 3TU.Datacentrum, perhaps in future Maps4Science)
2. Within Government (RWS, authentic registration)
3. Within the commercial sector (Fugro, added services)

7 Description of the proposed plan of work

The proposed project duration is 2 years: 2013-2014.

WP	WP title	Months	WP partners	Deliverables
WP1	Compile specifications/ requirements & Create Benchmarks (data, queries, etc.)	1-3	TU Delft OTB, TU Delft Library, Rijkswaterstaat, Fugro, Oracle	Requirements report; specifically described for all the different users; specifications of the datasets, queries and usage sceneries will be generated.
WP2	Design web-based interface, protocol between server and clients (WPCS supporting streaming data)	3-12	TU Delft OTB, TU Delft Library, Oracle, eScience engineer	Web Point Cloud Service for [1] selecting individual points from the point cloud; [2] selecting point cloud blocks; [3] transactional service to update point cloud (attributes). Proposal for web-standard to OGC/ISO.
WP3	Create facility, server side (hardware: storage, processing, communication)	8-12	TU Delft ICT, Oracle, eScience engineer	Operative configuration of facility
WP4	DBMS extension (with point cloud storage and analysis functionality)	11-13	TU Delft OTB, Oracle, eScience engineer	Research on DBMS processing; research report on the benefits of using state-of-the-art DBMS technology. Implementation of new PC data type (multi-resolution/LoD). Proposal for SQL-standard to OGC/ISO.
WP5	Design database schema, load data, tune (spatial clustering/indexing)	9-15	TU Delft OTB, TU Delft ICT, Oracle, eScience engineer	Database schema definitions (SQL DDL) for point cloud database including definition of primary storage structure (clustering), definition of (spatial) indexing, tuning of block point cloud size, and actual bulk loading 5 Tb of AHN2-data
WP6	Develop client side (pose queries, visualize results)	12-22	TU Delft OTB, TU Delft Library, Oracle, Fugro, eScience engineer	Client-side will support following: functionality: [1] multi-resolution rendering; [2] new update methods; [3] techniques for smooth transitioning between different LoD; [4] progressive streaming mechanism which allows low-bandwidth transfers; [5] investigate potential use of Graphic Processing Units
WP7	Execute Benchmark (incl. scaled up versions of data set)	a:14-16 b:22-24	TU Delft OTB, TU Delft ICT, Rijkswaterstaat, Fugro, Oracle	[1] testing of the point cloud service with datasets of various scales to test performance and the ability to scale (a: at SQL level, b. at web-services and visualization level); [2] test reports.
WP8	Investigate operational model for facility after the project	20-24	TU Delft OTB, TU Delft Library, Rijkswaterstaat, Fugro, Oracle	[1] report on key issues for long term availability; [2] market research report; [3] Operational model proposal for: academic, government, and commercial users.
WP9	Project management	1-24	TU Delft OTB	[1] Organize project meetings; [2] Project Implementation Document; [3] Time sheets; [4] budget overviews; [5] progress & end reports; [6] project communication internal and external (publications).

WP1: Compile specifications/requirements & Create benchmarks

In this work package the basic functionalities are identified that are needed in a point cloud spatial Database Management System (running on parallel hardware) to support the other work packages. This will be done by [1] literature research, [2] discussions with the project members, and [3] customer surveys. It will result in a report. As this work package precedes all of the other work packages of the project the generated report in this work package will define exactly what the other work packages need from the foreseen parallel hardware and the DBMS.

The requirements have a strong relationship with the planned benchmark. The benchmarks will give precise information on the requirements that must be complied by the parallel system and will specify the datasets (AHN2 for 'production' and optionally other laser scanning data for scalability testing and for proofing the generic character of the realized solution), the queries and the usage sceneries. The benchmark reflects the three main levels of point clouds use:

a. SQL level testing (DBMS);
b. web-services testing (including progressive communication); and
c. client side visualization and analysis.

The type of queries (selections and computations) in the benchmark should include (at all three levels) a (1) simple range/rectangle filters (of various sizes), and selections based on (2) points along a linear route/polyline; (3) of the attributes I (or in future also RGB values when included in AHN2, compare illustrations page 3 and 17); (4) multi-resolution/LoD; (5) slope orientation or steepness; (6) nearest neighbor; (7) point cloud density; (8) spatial with other table; e.g. polygons of buildings; (9) spatio-temporal selection queries; and (10) temporal differences computations and selection, etc. All these queries will be selections using 2D or 3D primitives, return exact answers (up to point level) or larger sets (e.g. filtered blocks containing answer), and limit answers by a. amount of data, or amount of elapse time. Both visualisation and analysis consequently rely heavily on the computation of surface normals (Mitra, 2004). This is a very demanding and time consuming task, which can be supported by smart indexing of the point cloud (Muja, 2009). Having indexed point clouds will be very beneficial to the end users (see WP5).

WP2: Design web-based interface, protocol between server and clients (WPCS supporting streaming data)

The foreseen point cloud data server stores the point cloud data and this data should be accessible using, preferably, an open standard. This standard does not exist yet, but in analogy with other Open Geospatial Consortium (OGC) standards for distributing 2D GIS maps, it is called Web Point Cloud Service (WPCS) in this project. In the project will be investigated whether the new WPCS can be made compatible with existing standards for file based laser data. The protocol service provides an interface for:

1. Selecting individual points from a point cloud (including attributes) within a 3D bounding box.
2. Selecting points with the right density on the fly, depending on bounding box size and viewing distance (near more points, far less points).
3. Transactional service to update point cloud (attributes such as classification, normal vector, etc.)

It should be noted that the broadband connection itself is the greatest bottleneck for such a service, since the data volumes are large. This can only be overcome by smart indexing of the point clouds, data compression and the application of LoDs. It is important that this will be standardized. A proposal for standardization will be made in this work package to ISO/TC211 and/or OGC. Similar to the current OGC web services, the WPCS data streams can be consumed by a program, currently identified as the client application (see WP6).

WP3: Create facility, server side (hardware: storage, processing, communication)
The described reference architecture in section 6.b of this proposal will be implemented in this work package. More details with respect to the actual instrumentation used will be described in section 8 of this proposal.

WP4: DBMS extension (with point cloud storage and analysis functionality)
The project aims at making a prototype implementation of a collection of database types plus operations for the handling of point cloud data. The new database types will integrate well with existing types for vector and raster data and will be designed to scale well within developed ICT infrastructure (WP3). As current implementations are often file based, research within this work package aims to investigate and show that a DBMS solution offers richer functionality. In addition to traditional features, such as data access control and multi-user concurrent access, a state-of-the-art DBMS with advanced spatial indexing can provide fast extraction of PC subsets of interest as needed by analysis and visualization. Such features are missing or very limited in the current file-based solutions. The design of a specialized point cloud index structure can built upon recent research (Muja, 2009). This new point cloud data type in the DBMS should support the concept of efficient clustering (in blocks), spatial indexing, support multi-resolution/LoD, enable compression, and provide functionality to realize the WPCS (of WP2).
The new data type in the DBMS should support all phases in the data lifecycle (i.e. geo-referencing, inserting/updating, change detection, modeling, visualization, application). Benefits of a well-designed geo-spatial database schema implemented using state-of-the-art DBMS technology will be researched and documented. A proposal for SQL-standardization will be made to ISO/TC211 and/or OGC.

WP5: Design database schema, load data, tune (spatial clustering/indexing)
This work package covers the actual work on setting-up and loading the database for the AHN2 data set (in the prepared ICT environment of WP3 and with the DBMS extended with point cloud support in WP4). This starts by providing the database schema definitions in SQL DDL (data definition language) for the AHN2 point cloud database including the definition of a primary storage structure (clustering), and a definition of (spatial) indexing, and other tuning parameters. This work package also includes the actual bulk loading 5 Tb of AHN2-data set after specifying the physical 'tuning' parameters based on research in WP4. And, if compression is selected, apply compression techniques (among other use 'local block coordinates'). Also, after or during loading the multi-resolution, a data pyramid will be created. Finally, after loading, indices are computed and statistics for the database query optimizer are collected. The design will be based on the point cloud data type developed in WP4.

WP6: Develop client side (pose queries, visualize results)
The first main challenge in this work package is to display as much information of the data as required for the task or the desired quality of visualization. The point cloud data on the servers is (and will be) several orders of magnitude larger than what fits in main memory and what can be rendered fluently on the display client. Therefore, only a subset of the data can be loaded into memory and displayed interactively. A transformation (preprocessing) of the point cloud into specialized (spatial) data structures for visualization is required. To achieve high-quality rendering, a very high number of points have to be displayed to visually create the illusion of closed surfaces. Therefore a multi-resolution data structure will be developed that is specifically tailored for rendering purposes. The problem of data consistency (e.g. data duplication, stale data, caching) between the spatial database and the optimized data structure(s) for rendering and editing is challenging and will be addressed in close cooperation with WP5.

Techniques for transitioning between different LoD will be designed. Once this data structure is available, scan attributes and viewing behavior of users are taken into account to perform view-dependent rendering. The data needed for rendering will be further reduced by detection of visually salient points or parts of the scene. To enable rendering on remote display clients data transfer aspects will be addressed, such as rendering latency and bandwidth limitations. A progressive streaming mechanism will be devised which allows low-bandwidth transfers, for use on web-clients or over mobile networks. One very specific aspect is to exploit the Graphic Processing Units/GPU's (at client side) given the spatial nature of most GIS related computations. Fugro will provide an example web-service based client implementation (WPCS) in its existing product Flaim (http://www.fugroearthdata.com/servicessubcat.php?subcat=lidar-mapping). This is a product to do currently only file base operations on LiDAR data. This implementation will serve as a demonstrator to show how multiple client programs can use the same data service.

WP7: Execute Benchmark (incl. scaled up versions of data set)
In this work package the new point cloud service will be tested by running the queries defined in work package 1. The tests will be run on the AHN2 dataset. In addition, the same queries will be tested on a larger dataset (factor 10 to 100) potentially by 'multiplying' and translating copies of AHN2. This will give information on the performance of the system in general and the ability to scale. The benchmark (test) results will be documented in a report.

WP8: Investigate operational model for facility after the project
In this work package research will be done on the identification of key issues for long term availability such as: ownership/licensing, funding, cost factors, data storage and archiving. Market research (suppliers, business models) will be executed that focus on the foreseen categories of users: academia, government and commercial/industry.
The project results will contribute to developments in all disciplines that are using point cloud datasets. The design and benchmark results will be shared in a publication and the developed code within the project will be made openly available. The services developed during the project and the infrastructure will be managed by the present consortium and a sustainable model for funding for data management and further development will be developed in this WP (see also section 6.c).
After the project, the scale of usage will be broadened to serve more users for these point cloud datasets. In the near future, the AHN2 data will be distributed for free due to the open data policy of the Dutch Government, so the range of users will be extended. The AHN2 dataset is expected to be completed for the Netherlands in March 2013. In 2014, the AHN2 will be gradually updated, with possibly higher point density and additional attributes (like RGB/colour values).

WP9: Project management
Project management will be lead by the project leader Prof. dr.ir. Peter van Oosterom, and executed by a management team consisting of one representative per partner. The objective of this work package is to provide coordination for the project, align the activities and deliverables across the work packages, ensure quality control of project deliverables, risk management, administrative and financial tasks connected with the activities of the project (controlling the budget), facilitate communications between project participants (and other relevant stakeholders) through regular telephone, conferences, email and project meetings as necessary.

8 Expected Use of Instrumentation – Relevance to eScience

Relational databases have developed sophisticated extensions to support geographic data (e.g. Oracle Spatial, with spatial indexing, many spatial functions). For most massive datasets this works quite well (including all sorts of raster and vector data). However, for point cloud data, it is very hard to deliver the services/performance as expected by users. One of the key issues is being scalable (more data, higher density point clouds, bigger areas than just NL, more attributes per point/return signals, combining point clouds with other massive data sets; e.g. optical images, collecting temporal time-series, but also more users and more heavy queries/selection/processing tasks). It is not difficult to imagine realistic scenarios with near future use of hundreds of Tb data (or even Pb range). So for high scalability test with 100TB data set we consider temporary usage of Oracle Cloud DB services. Within the project the main ambition is being able to handle AHN2 data very well ('just a few Tbs') and make a 'permanent' service for this. It is very important that the solution is generic and not just focusing on a single specific data type (point clouds) and query type. Therefore, the object-relational DBMS forms the point of departure. To be able to realize the abovementioned solution a suite of Oracle products will be used, since they offer momentarily the most spatial functionality (see also section 6.b). The focus will be primarily on the Oracle Database and specifically the new (improved) Point Cloud capabilities of Oracle Spatial and Graph, but also optional components which can help leveraging the application platform reliability, availability, scalability and performance requirements.

Since Oracle Spatial and Graph is a native function of the Oracle Database, the computing foundation characteristics are fully applicable, thus providing system and storage virtualization to provide a highly scalable infrastructure in terms of performance and data size. The basic requirements to setup the demanded infrastructure for the Oracle Database are as follows:

- Modern hardware, based on i86 processor technology
- Storage array (SAN or NAS based)
- Oracle Enterprise Linux (RedHat based) or Solaris; this is not a business requirement, but a project requirement to stay in pace with Oracle Product Development (patching)
- Oracle Database (Enterprise Edition 11gR2 or 12c Beta)
- Oracle Spatial
- Oracle Real Application Cluster & Oracle Automatic Storage Management; for system en storage virtualization
- Oracle Partitioning/Oracle Advanced Compression; for data partitioning and compression

This setup is identical to the Oracle Database Machine (1), providing additional performance and storage optimizations due to the optimized hardware and software integration. To persist the data from Hadoop, the Hadoop loader for Oracle Data Integrator (ETL) is capable of offloading the data to the Oracle Database. The Oracle Big Data Appliance (2) combines Hadoop with Oracle Data Integrator functionality to combine the processing and storage capabilities in conjunction with an In Memory (NoSQL) Database. Lastly, Oracle Coherence, a function of Oracle's Java Application Platform (3) allows creating a Data Grid based on Objects shared over multiple nodes in a network. Oracle Exalogic combines this with hardware. Components (1) and (2) are Cloud ready (with or without hardware) and respectively as Database as a Service or Java as a Service/Middleware as a Service.

Oracle will provide the TU Delft with temporary software licenses and hardware to assist in testing and development (see also Commitment Letters in active hyperlink below table 'Composition of the research team'). No software has to be purchased. In the budget €45k is proposed to fund a permanent facility for storage.

9a Literature

- ASTM, 2011. American Society for Testing and Materials (ASTM) E2807 - Standard Specification for 3D Imaging Data Exchange, Version 1.0, 2011 (http://www.astm.org/Standards/E2807.htm).
- Craglia, M., De Bie, K., Jackson, D., Pesaresi, M., Remetey-Fülöpp, G., Wang, C., Annoni, A., Ling, B., Campbell, F., Ehlers, M., Genderen, J.V., Goodchild, M., Guo, H., Lewis, A., Simpson, R., Skidmore, A. & Woodgate, P., 2012. Digital Earth 2020: towards the vision for the next decade. International Journal of Digital Earth, 5 (1), 4-21.
- M. Isenburg, 2011. LASzip: lossless compression of LiDAR data. European Lidar Mapping Forum, Salzburg, 2011.
- Kroes, N., 2012, Giving Europe's scientists the tools to deliver, speech at International Conference on Research Infrastructures (ICRI 2012), Copenhagen, 23 March 2012 (http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/12/218&format=HTML&aged=0&language=EN&guiLanguage=en)
- N.J. Mitra, A. Nguyen, L. Guibas, 2004. Estimating Surface Normals in Noisy Point Cloud Data, International Journal of Computational Geometry & Applications.
- M. Muja, D. Lowe, 2009. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration, International Conference on Computer Vision Theory and Applications (VISAPP'09).
- R. Schnabel, R. Wahl, R. Klein, 2007. Efficient RANSAC for Point-Cloud Shape Detection, Computer Graphics Forum.
- M.E. Wijga-Hoefsloot, 2012. Point Clouds in a Database, Data Management within an Engineering Company. TU Delft Master Thesis, in preparation, September 2012 (Note: currently working for Fugro).

9b Key publications of the Research team

1. M.P. Kodde, P. Bakhuizen, 2011. Meten, karteren of beheren - Landmeetkunde in 2015 draait om data, software en integratie, Geo-Info 2011, nummer 1, pages 8-9.
2. P.J.M. van Oosterom, C.W. Quak and T.P.M. Tijssen, 2001. Testing current DBMS products with real spatial data. In: E.M. Fendel, K. Jones, R. Laurini and M. Rumor (Eds.); Proceedings of UDMS '02, 23rd Urban Data Management Symposium, Prague, 2002, pp. VII.1-VII.18
3. P.J.M. van Oosterom, M.G. Vosselman, Th.A.G.P. van Dijk and M. Uitentuis (Editors), 2010. Management of massive point cloud data: wet and dry, Royal Netherlands Academy of Arts and Sciences - Netherlands Geodetic Commission (KNAW-NCG), Nr 49, Delft, 2010, 104 pages.
4. E. Verbree, T.P.M Tijssen, H. Ledoux and S. Zlatanova, 2012. 3D Pilot - Eindrapport werkgroep 3D Testbed (in Dutch). NCG-KNAW serie, Nederlandse Commissie voor Geodesie, nr. 54, Delft, 2012, 31 p. (http://www.ncg.knaw.nl/Publicaties/Groen/54Verbree3DPilotTestbed.html).
5. N. van der Zon, 2012, Kwaliteitsdocument AHN, Actueel Hoogtebestand Nederland rapport, datum: 5-6-2012, Versie:1.2, (http://www.ahn.nl/$mab/page/)