## Parámetros del *benchmarking*.

**Objetivo:** Plantear las métricas a usar y la forma de cómo se llevará a cabo el *benchmarking.*

**Resultados esperados:** Documento con una lista de parámetros que serán usados en el *benchmarking*.

**Definiciones**

*Benchmarking*: Proceso metódico y continuo para evaluar servicios, productos y/o procesamientos de trabajo de organizaciones o empresas las cuales son bien reconocidas por llevar a cabo las mejores prácticas con el objetivo que mejorar o llevar a cabo estudios de referencia [1].

Plan de ejecución: Secuencia de operaciones en sistemas de administración de bases de datos para llevar cabo las sentencias que se requieran [2].

**Desarrollo**

En el presente documento se explican las bases del *benchmarking* que se realizará una vez completado el módulo de consultas federadas para el software Apache Marmotta.

Michael Shmidt y sus colegas [3] en su *paper FedBench: A Benchmark Suite for Federated Semantic Data Query Processing* proponen un *benchmarking* que mida las capacidades de los gestores de consultas federadas. Fue elaborado basándose en otros *benchmarking* populares y citados*,* los cuales son *SP^2Bench* [4] y *Berlín SPARQL Benchmark* [5]*,* cuyo objetivo fueron evaluar consultas centralizadas. Este será el documento en el se basará para construir el *benchmarking* del actual proyecto.

Tal como lo establece Michael Schmidt y su equipo, no existe una evaluación que cubra todos los aspectos de un sistema de consultas. Sin embargo, ellos abordan la heterogeneidad a nivel de datos el cual, es uno de los retos que los sistemas de consultas tienen que abordar.

Heterogeneidad a nivel de datos se identifican en los siguientes aspectos

* Distribución física: Cómo se comportan los sistemas de consultas frente a *datasets* cuya ubicación física son distintas.
* Interfaz de acceso a datos: El acceso a los datos pueden ser repositorios nativos, *SPARQL endpoints* y datos en la nube del *Linked Data*.
* Existencia de fuentes de datos: Este aspecto solo es posible llevarse a cabo en la nube del *Linked Data* ya que a priori no se puede averiguar si la nube de datos contiene o no las fuentes de datos solicitadas en la consulta. De manera local, si se tiene el conocimiento sobre los campos que hay en el repositorio.
* Estadísticas de datos: De manera local, se puede conocer información de los repositorios mediante sus histogramas sobre propiedades, cantidad de datos y su distribución de datos. Sin embargo, cuando los repositorios no son locales, puede haber poca o nula información acera de ellos.

Aunado a lo anterior, también existen retos a nivel de consultas. Michael y sus colegas identificaron los siguientes aspectos a cubrir en la evaluación de consultas.

* Lenguaje de consulta: La expresividad entre lenguajes de consulta puede variar. Unos son simples, mientras que otros basan su sistemas en lenguajes de consultas muy específicos.
* Completitud de datos: Existen sistemas que pueden lidiar con una gran cantidad de información, pero también hay otros que no. Cabe decir que en la nube del *Linked Data* no es posible determinar la completitud de datos ya que un *triple store* puede desembocar en otros *triple store*, dando como resultado una interminable búsqueda de datos o quizá, un *triple store* no está referenciado correctamente a otros *triple store* resultando en resultados truncados.
* Lista de clasificación: Los sistemas pueden tener establecido una lista de mejores consultas basadas en sus propias métricas.

*FedBench* está basado en 3 componentes los cuales, pueden ser modificados, extendidos y personalizado para cubrir un escenario en específico. Los componentes son

* Múltiples *datasets*: Elaboración de *datasets* y estadísticas. Para este proyecto, los *datasets* serán sobre datos geoespaciales.
* Múltiples conjuntos de consultas: Elaboración de las consultas.
* *Framework* de evaluación

La continuación del *paper* es una profundización de los 3 componentes mencionados, pero con sus propios datos y consultas. Tal y como el mismo *paper* lo propone, para este proyecto terminal se adaptará los componentes para llevarlo a cabio con el software de Apache Marmotta una vez implementado el módulo.

Se determinó que el *paper* anterior no era suficiente para delimitar el *benchmarking* del proyecto actual por lo que se buscó otro.

Gabriela Montoya y su equipo en España desarrollaron un *paper* que cubriera las limitaciones del *benchmarking* que Michael había desarrollado. El *paper* B*enchmarking Federated SPARQL Query Engines: Are Existing Testbeds Enough?* [6]especifica más sobre *FedBench* y sobre los componentes que impactan el rendimiento de sistemas de consultas federadas.

Con base a Montoya y su equipo, existen 2 variables que influyen en el sistemas de consultas federadas y son

* Independientes: Son aquellas características que debe de ser especificadas con el fin de asegurar que sean replicables los escenarios de evaluación. Las variables independientes que se proponen son
  + Consulta: Variable que definen las consultas en términos de su estructura, evaluación y expresividad del lenguaje. Está basada en 3 aspectos
    - Forma del plan de ejecución: La forma en que los planes de ejecución se generan en los sistemas pueden variar, por lo que tiene un importante impacto en las 3 variables dependientes.
    - Número de patrones de tripletas básicas en la consulta: Este aspecto afecta directamente al plan de ejecución debido a que depende del número de tripletas que se recuperan en función de las tripletas especificadas en la consulta.
    - Instancias y posición en las tripletas: Este aspecto está relacionado a cómo están relacionados el sujeto, el objeto y el predicado en las tripletas. Puede tener un gran impacto en la evaluación este aspecto ya que, por ejemplo, el predicado de una tripleta puede no estar instanciado dando lugar a que diversos *endpoints* respondan cuando una consulta requiera información de dicho predicado.
  + Datos: Esta variable está compuesta por 4 elementos
    - Tamaño del RDF *dataset*.
    - Características estructurales del *dataset* las cuales se refueren al número de predicados, objetos y sujetos. Estas características influyen en el número de tripletas que son transferidas y por consecuencia, el tiempo del plan de ejecución total.
    - Tipo de partición: Se refiere a cómo el *dataset* está fragmentado. La partición puede ser horizontal y vertical. Dependiendo del tipo de partición influye en las 3 variables dependientes ya que la cantidad de información puede ser menor, al tener menor cantidad de filas, o parcial al solo considerar ciertas propiedades.
    - Distribución de datos: El cómo las particiones están distribuidas en los *endpoints* también afecta las tres variables dependientes. Pueden estar completamente distribuido, centralizado o mixto.
  + Plataforma: Esta variable agrupa a otras relacionadas a la infraestructura de cómputo durante la evaluación. Los elementos que componen esta variable son.
    - Memoria RAM.
    - Número de procesadores.
    - Administración de memoria caché.
  + *Endpoint*: En esta variable se refiere a los elementos asociados al número y capacidades de *SPARQL endpoints* usados en la evaluación. Los elementos son
    - Número de *SPARQL endpoints* a visitar por la consulta cuando esta sea llevada a cabo.
    - Tipo de *SPARQL endpoints* usados en la evaluación.
    - Distribución de las transferencias de datos el cual es la distribución de tiempos de transmisión hechos por los *SPARQL endpoints.*
    - Latencia de red: Retraso en el envío de paquetes a través de la red.
    - Retraso inicial del *SPARQL endpoint*.
* Dependientes: Son las características que están influenciadas por las variables independientes. Las variables dependientes son las que serán medidas en la evaluación. Las variables dependientes que se proponen son
  + Tiempo de selección del *endpoint:* Tiempo transcurrido entre la realización de la consulta y la generación de la consulta federada, basado en el protocolo *SPARQL* 1.1, con los *endpoint* donde se llevarán a cabo las subconsultas.
  + Tiempo de ejecución: Esta variables se comprende de 3 elementos
    - Tiempo entre la realización de la consulta y la primer respuesta.
    - Distribución de tiempo en la recepción de respuestas de las consultas.
    - Tiempo total de ejecución.
  + Completitud de respuestas: Cantidad de respuestas recibidas comparados a los datos disponibles en los *endpoints* seleccionados.

Así como se dio la definición al inicio, el *benchmarking* permitirá describir el módulo de consultas federadas y hacer una comparación respecto a otros sistemas similares. Opcionalmente, ya que no es el objetivo de la actividad, se podría optimizar el módulo de consultas federadas en el *software* Apache Marmotta utilizando los resultados obtenidos del *benchmarking*

La tabla 1 muestra la relación existente entre las variables independientes y las dependientes.

Tabla Relación existente entre variables dependientes e independientes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables independientes | | Variables dependientes | | |
| Tiempo de selección del *SPARQL endpoint* | Tiempo de ejecución | Completitud de la respuesta |
| Consulta | Forma del plan de ejecución | Si | Si | Si |
| Número de patrones de tripletas básicas en la consulta | Si | Si | Si |
| Instancias y posición en las tripletas | Si | Si | No |
| Datos | Tamaño del RDF *dataset* | No | Si | No |
| Características estructurales del *dataset* | No | Si | No |
| Tipo de partición | Si | Si | Si |
| Distribución de datos | Si | Si | Si |
| Plataforma | Memoria RAM | Si | Si | No |
| Número de procesadores | Si | Si | No |
| Administración de memoria caché | Si | Si | No |
| *SPARQL endpoint* | Número de *SPARQL endpoints* | Si | Si | Si |
| Tipo de *SPARQL endpoints* | Si | Si | No |
| Distribución de las transferencias | Si | Si | Si |
| Latencia de red | Si | Si | Si |
| Retraso inicial del *SPARQL endpoint* | Si | Si | No |

En resumen, los 3 parámetros a evaluar, en segundos los 2 primeros y en número de tripletas el tercero, en el *benchmarking*son

* Tiempo de selección del *SPARQL endpoint*
* Tiempo de ejecución.
* Completitud de respuesta.

Cabe decir que el *benchmarking* a realizar en el proyecto terminal 2, estará contextualizado en datos geoespaciales por lo que las consultas [7] serán modificadas y adaptadas, tal y como *FedBench* lo sugiere.

Una lista de *triple store* a consultar pueden ser los mostrados en la tabla 2, aunque no son definitivos.

Tabla Triple stores y su cantidad de tripletas disponibles

|  |  |
| --- | --- |
| Sitio Web | Total tripletas |
| http://www.geonames.org/ontology/ | 93,896,732 |
| http://lod.taxonconcept.org/ | 34,394,994 |
| http://eunis.eea.europa.eu | 20,229,105 |
| http://data.fishesoftexas.org | 528,897 |
| http://lod.geospecies.org/ | 2,163,922 |
| http://ja.dbpedia.org/ | 70,790,371 |
| https://www.geospatialworld.net/ | 27,017 |
| http://linkedgeodata.org/ | 3,000,000,000 |
| http://www.best-project.nl | 1,931,860 |
| http://webenemasuno.linkeddata.es/ | 9,262,184 |
| http://geo.linkeddata.es/ | 21,564,199 |
| http://www.rdfabout.com/demo/census/ | 1,002,848,918 |
| http://www4.wiwiss.fu-berlin.de/eurostat/ | 8,850 |
| http://www4.wiwiss.fu-berlin.de/factbook/ | 38,640 |
| http://data.ordnancesurvey.co.uk/ | 2,788,872 |
| http://statistics.data.gov.uk/ | 343,733 |
| https://www.pdok.nl/introductie?articleid=1948911 | 1,000,000,000 |
| http://mapasinteractivos.didactalia.net/comunidad/mapasflashinteractivos | 90,226 |
| http://thesaurus.iia.cnr.it/index.php/vocabularies/earth | 133,315 |
| http://purl.org/NET/ThISTWebPage | 396,885 |
| http://www.morelab.deusto.es | 3,561 |
| http://transparency.270a.info/ | 41,592 |
| http://worldbank.270a.info/ | 175,000,000 |
| http://id.ecs.soton.ac.uk/docs/ | 50,000 |
| http://eprints.ecs.soton.ac.uk/ | 1,000,000 |
| http://telegraphis.net/data/ | 15,257 |
| http://nuts.psi.enakting.org/ | 40,000 |
| http://nuts.geovocab.org/ | 300,000 |
| http://www.openmobilenetwork.org | 19,048,586 |
| https://lod-cloud.net/dataset/linkedcrowdsourceddata | 17,043 |
| http://lobid.org/resource | 702,686,189 |
| http://lobid.org/organisation | 776,735 |
| http://purl.org/NET/marccodes/ | 8,816 |
| http://purl.org/smartlink | 2,829 |
| http://education.data.gov.uk/ | 6,619,847 |
| http://ec.europa.eu/eurostat/ramon/rdfdata/ | 68,000 |
| http://wiki.knoesis.org/index.php/SSW\_Datasets | 1,730,284,735 |
| http://www.opencalais.com/ | 4,500,000 |
| http://openlylocal.com/ | 10,000,000 |
| http://theviewfrom.org/ | 368 |
| http://transport.data.gov.uk/ | 329,527,661 |
| http://www.uk-postcodes.com/ | 27,137,408 |
| https://lod-cloud.net/dataset/geoecuador | 1,278,288 |
| http://www.kanzaki.com/works/2011/stat/ra/ | 1,648,890 |
| http://greek-lod.math.auth.gr/fire-brigade/ | 4,134,837 |
| http://greek-lod.math.auth.gr/police/ | 145,368 |
| http://www.linkedopenservices.org/services/geo/SpatialResources/point/ICAO/ | 75,184 |
| http://eurostat.linked-statistics.org/ | 8,000,000,000 |
| http://geowordnet.semanticmatching.org/ | 53,390,969 |
| http://kent.zpr.fer.hr:8080/educationalProgram/ | 2,373,863 |
| http://datos.fundacionctic.org/en/ | 535,230 |
| http://seobook.blog.com | 40,798 |
| http://data-gov.ie/ | 12,390,973 |
| http://www.fao.org/countryprofiles/geoinfo.asp?lang=en | 22,495 |
| http://ontologycentral.com/2009/01/eurostat/ | 40,000,000 |
| http://data.lenka.no/ | 8,399 |
| http://airports.dataincubator.org/ | 748,363 |
| http://www.josemalvarez.es/web/2011/11/01/nomenclator-asturias-2010/ | 4,508,050 |
| http://www.josemalvarez.es/web/2011/11/16/product-scheme-classifications/ | 1,842,053 |
| http://kasabi.com/dataset/yahoo-geoplanet | 49,734,022 |
| https://lod-cloud.net/dataset/ww1lod | 40,160 |
| http://greek-lod.math.auth.gr/kalikratis/ | 31,068 |
| http://data.oceandrilling.org/linkeddata.html | 24,489,366 |
| http://data.oceandrilling.org/codices/ | 24,489,366 |
| http://ecowlim.tfri.gov.tw/lode | 23,015,257 |
| http://www.ldf.fi/dataset/warsa | 12,677,886 |
| http://www.linkedopendata.it/datasets/grrp | 153,935 |
| http://smcjournals.dataincubator.org | 15,801 |
| http://climb.dataincubator.org/ | 174,120 |
| http://pleiades.stoa.org/ | 2,600,000 |
| http://chronos.org/janusAmp/ | 24,489,366 |
| http://aemet.linkeddata.es/ | 62,932,032 |
| http://environment.data.gov.uk/lab | 8,735,962 |
| http://metoffice.dataincubator.org | 26,000,000 |
| http://ieeevis.tw.rpi.edu | 19,935,340 |
| http://kasabi.com/dataset/renewable-energy-generators | 99,917 |
| http://onki.fi/en/browser/overview/kunnat | 11,380 |
| http://eurobarometer.publicdata.eu | 1,193,494 |
| http://my-family-lineage.com/wiki/Main\_Page | 3,000,000 |
| http://opendatacommunities.org/datasets/imd-score-2010 | 194,892 |
| http://riese.joanneum.at/data/ | 5,000,000 |

Los datos de la tabla 2 se extrajeron del portal web de la nube del *Linked Data[[1]](#footnote-1)* usando un *script* en Python para extraer los sitios web que tengan datos geoespaciales. El script es el que se muestran en la figura 1 y 2. El código puede visitarlo en el GitHub[[2]](#footnote-2) del estudiante.

Fig. Web scrapper para la nube de Linked Data

Imagen que contiene captura de pantalla

Descripción generada automáticamente

Fig. Web scrapper para la nube de Linked Data - continuación

# Referencias

|  |  |
| --- | --- |
| [1] | UNAM, «Benchmarking,» UNAM, 23 Noviembre 2018. [En línea]. Available: http://depa.fquim.unam.mx/amyd/archivero/TRABAJO7\_2848.pdf. [Último acceso: 19 Octubre 2019]. |
| [2] | D. Tow, SQL Tuning: Generating Optimal Execution Plans, California: O'Really Media, In, 2003. |
| [3] | M. Shmidt, O. Görlitz, P. Haase, G. Ladwig, A. Shwarte y T. Tran, «FedBench: A Benchmark Suite for Federated Semantic Data Query Processing,» de *International Semantic Web Conference*, Berlín, 2011. |
| [4] | M. Schmidt, T. Hornung, G. Lausen y C. Pinkel, «SP^2 Bench: a SPARQL performance benchmark,» de *2009 IEEE 25th International Conference on Data Engineering*, Freiburg, Alemania, 2009. |
| [5] | C. Bizer y A. Schultz, «The berlin sparql benchmark,» *International Journal on Semantic Web and Information Systems,* vol. 5, nº 2, pp. 1-24, 2009. |
| [6] | G. Montoya, M. E. Vidal, O. Corcho, E. Ruckhaus y C. Buil-Aranda, «Benchmarking Federated SPARQL Query Engines: Are Existing Testbeds Enough?,» de *International Semantic Web Conferernce*, Berlín, 2012. |
| [7] | C. Buil-Aranda y O. Corcho García, Federated Query Processing for the, Madrid: IOS Press, 2012. |

1. [https://lod-cloud.net](https://lod-cloud.net/) [↑](#footnote-ref-1)
2. <https://github.com/Osw1997/PT1-scripts.git> [↑](#footnote-ref-2)