بهبود مدل مفهومی CLAaaS برای تحلیل داده‌های عظیم در ابر

لیلا محمدی نافچی *[1]، محمدعلی منتظری[2]، علی آهون‌منش[3]

[1] دانشجوی مقطع کارشناسی ارشد، موسسه آموزش عالی صنعتی فولاد، فولادشهر،
Lm.eng@live.com
[2] استادیار گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی اصفهان،
montazeri@cc.iut.ac.ir
[3] استاد، رئیس مؤسسه صنعتی فولاد، فولادشهر،
aliahoon@cc.iut.ac.ir

چکیده

تحلیل داده اهمیت خود را در کشف دانش و پشتیبانی از تصمیم‌گیری در حوزه‌های کاربردی و اطلاعاتی متفاوت به اثبات رسانیده است. تحلیل داده‌های عظیم چالشی جدی در خصوص منابع سخت‌افزاری و نرم‌افزاری مورد نیاز را مطرح نموده است. امروزه تکنولوژی ابر راه‌حلی امید بخش برای این چالش‌ها به وسیله امکان تأمین منابع محاسباتی به صورت مقیاس‌پذیر و فراگیر پیشنهاد کرده است. با این حال چالش‌های بسیاری باقی مانده است، مانند در دسترس بودن نرم‌افزار تحلیلی مورد نیاز برای حوزه‌های کاربردی گوناگون، تخمین و به اشتراک‌گذاری منابع مورد نیاز برای کار تحلیل یا جریان کاری، مدیریت داده در ابر، و طراحی، ارزیابی و اجرای جریان‌های کاری تحلیل، که باید به اینها پرداخته شود. در این مقاله روشی برای بهبود معماری مفهومی تحلیل مبتنی بر ابر به عنوان CLAaaS[1] که به وسیلهٔ آن روند تحلیل با توجه به این که یک معماری همه جانبه برای تمامی زمینه‌ها می‌باشد، تسریع بخشیده می‌شود، معرفی شده است.

کلمات کلیدی

تحلیل داده‌های عظیم، ابر، CLAaaS، هادوپ[2]

* اصفهان- فولادشهر- مؤسسه آموزش عالی صنعتی فولاد – کد پستی ۸۴۹۱۶۶۳۷۶۳

[1] Cloud-based Analytics-as-a-Service
[2] Hadoop

۱- مقدمه

امروزه در جهان کسب‌وکار بسیار پیچیده، سازمان‌ها باید راه‌های مبتکرانه برای متفاوت بودن با رقبا را به وسیله متحد شدن بیشتر، صحت بیشتر، وفق دادن، هماهنگی و چابک‌تر شدن پیدا کنند. آن‌ها نیاز دارند که قادر به پاسخ-گویی به نیازهای کسب‌وکار و تغییر سریع باشند. آن‌ها باید متوجه باشند که صاحب چه داده‌هایی هستند و چگونه متفاوت از دیگران از آن‌ها استفاده کنند. داده‌ها و اطلاعات در حال تبدیل شدن به دارایی اصلی برای بسیاری از سازمان‌ها می‌باشند[1]. تجزیه و تحلیل داده اثر بالقوه خود را در فراهم کردن پشتیبان تصمیم‌گیری در بخش‌های مالی، مدیریتی و علمی با ایجاد قابلیت محاسبات پیچیده برای بدست آوردن دانش، بینش و نتایج تجربی برای اکتشافات علمی اثبات کرده است[2]. واژه تحلیل حوزه وسیعی را که شامل تکنیک‌های تحلیل و پشتیبان تصمیم‌گیری است، ارائه می‌کند. برخی از کارهای تحلیلی قطعی‌اند و دارای خروجی مشخصی می‌باشند که این خروجی‌ها به سیستم پشتیبان تصمیم‌گیری داده می‌شوند ولی برخی دیگر بیشتر حالت اکتشافی دارند به طوری که نتیجه نهایی ممکن است مفید باشد یا نباشد و معمولاً نیاز است که چندین بار مورد بازبینی قرار بگیرند تا نتیجه قطعی حاصل شود. برنامه‌های تحلیل اکتشافی اصولاً پویا هستند. یک فرآیند تحلیل داده که جریان‌کاری تحلیل نامیده می‌شود[3,4]، به طور کلی شامل وظایفی چون تجمیع و پاکسازی داده و یا کارهای اکتشافی مثل تعریف و اجرای مدل‌های یادگیری ماشین یا پرسش‌های تحلیلی ساده می‌باشد[۲].

چالش‌هایی نیز در رابطه با داده‌ها برای سازمان‌ها وجود دارد. برای مثال چالش مدیریت میزان زیاد داده‌ها (داده‌های عظیم) که به خاطر انباره‌های کوچک‌تر و تکامل تجهیزات جمع‌آوری اطلاعات و داده‌های دیجیتال مثل گوشی‌های همراه، لب‌تاپ‌ها و سنسورها به طور فزاینده در حال بزرگتر شدن می‌باشند[۱]. در نتیجه میزان داده‌های مورد نیاز برای تحلیل با نرخی نمایی در حال رشد می‌باشد و متخصصان تکنولوژی تحلیل داده مانند داده‌کاوی و یادگیری ماشین دانش کافی در مورد این که کدام یک از داده‌ها نیاز است که تحلیل شوند، ندارند[۲]. در سال‌های اخیر داده‌های عظیم به موضوعی فراگیر در کسب‌وکار، علوم کامپیوتر، مطالعات اطلاعات، آمار و بسیاری از زمینه‌ها گشته است[۵]. بر اساس تعریف گروه گارتنر، داده‌های عظیم به طور کلی به حجم بالا، سرعت بالای تولید داده و تنوع بالای سرمایه‌های اطلاعاتی اطلاق می‌شود که برای پردازش مبتکرانه و مقرون به صرفه به منظور بهبود ادراک و تصمیم‌گیری مورد استفاده قرار می‌گیرند[۶]. تفکر مبتنی بر سرویس یکی از الگوهای در حال رشد در فناوری‌اطلاعات می‌باشد که با بسیاری از سیستم-های دیگر نظیر حسابداری، مالی و عملیاتی در ارتباط می‌باشد[۷]. امروزه تکنولوژی ابر راه‌حلی امید بخش برای این چالش‌ها به وسیله امکان تأمین منابع محاسباتی به صورت مقیاس‌پذیر و فراگیر پیشنهاد کرده است. با این حال چالش‌های بسیاری باقی مانده است، مانند در دسترس بودن نرم‌افزار تحلیلی مورد نیاز برای حوزه‌های کاربردی گوناگون، تخمین و به اشتراک-گذاری منابع مورد نیاز برای کار تحلیل یا جریان کاری، مدیریت داده در ابر، و طراحی، ارزیابی و اجرای جریان‌های کاری تحلیل، که باید به اینها پرداخته شود[۲].

در این زمینه دلن و همکارش، سیستم‌های پشتیبان تصمیم‌گیری مبتنی بر سرویس (DSS در ابر) را مطرح کرده و یک مدل مفهومی که از ارزیابی تئوری و روش‌های مدیریت سیستم‌های پشتیبان تصمیم‌گیری مبتنی بر سرویس (SODSS) پشتیبانی می‌کند، ارائه داده‌اند[۱]. بخشی از این مقاله نگاهی به چگونگی تحویل محیط DSS مبتنی بر محصول دارد، و فرصت‌ها و چالش‌های مهندسی DSS مبتنی بر سرویس در ابر را تشریح می‌کند. زمانی‌که داده، اطلاعات و تحلیل به عنوان سرویس تعریف می‌شوند، دیده می‌شود که مکانیزم‌های اندازه‌گیری قدیمی که اساساً زمان‌بر و هزینه‌بر می-باشند، به خوبی کار نمی‌کنند. سازمان‌ها نیاز دارند که ارزش سطح سرویس و کیفیت را به علاوهٔ هزینه و مدت زمان سرویس تحویل داده شده، در نظر بگیرند. DSS در ابر مقیاس، چشم‌انداز و سرعت اقتصاد را توانمند می‌سازد. این مقاله متشکل از دانش جدید در علم سرویس به وسیله اتصال جنبه‌های استراتژی فناوری‌اطلاعات به جنبه‌های علم طراحی و پایگاه داده برای مخاطبان وسیع‌تر، می‌باشد[۱]. زالکرناین و همکارانش نیز برای مقابله با چالش داده‌های عظیم، یک طبقه‌بندی برای سیستم‌های جریان کاری تحلیل برای مشخص کردن ویژگی‌های مهم در سیستم‌های موجود ارائه داده‌اند. در این مقاله بر اساس طبقه‌بندی و مطالعه بر روی سیستم‌ها و نرم‌افزارهای تحلیلی موجود، یک معماری مفهومی از تحلیل مبتنی بر ابر به عنوان سرویس (CLAaaS)، بستری برای ایجاد سرویس تحلیل داده‌های عظیم در ابر پیشنهاد شده است. در این تحقیق ویژگی‌های مهم برای CLAaaS به عنوان یک سیستم تأمین سرویس مثل کمک و سفارشی‌سازی برای حوزه و کاربرد خاص، همکاری، معماری ماژولار برای توسعه مقیاس‌پذیر و توافق سطح سرویس (SLA) مطرح شده است[۲].

در کنار چالش مطرح شده به وسیله رشد سریع میزان داده‌ها، این فرصت بسیار بزرگی برای جهانی است که در حال دیجیتالی شدن بیشتر و بیشتر می-باشد، که اجازه تجمیع و تحلیل داده در زمینه‌ای خاص را می‌دهد[۱]. یکی از زمینه‌هایی که تحلیل داده‌های عظیم در آن مطرح می‌باشد و با چالش مربوط به آن مواجه است، علم هواشناسی می‌باشد. اشنیز[۳] و همکارانش، در مقاله خودشان به چالش‌های داده‌های عظیم در علم هواشناسی پرداخته‌اند، حرکت این مقاله به سمت مفهوم تحلیل آب و هوا به عنوان یک سرویس (CAaaS) می‌باشد. این مقاله به دلیل دانش کسب شده در تعامل با داده‌های عظیمی که در نهایت منجر به تولید فواید اجتماعی می‌شود، بر روی تجزیه و تحلیل تمرکز کرده است. اشنیز و همکارانش بر روی CAaaS تمرکز کرده‌اند زیرا بر این باورند که این موضوع راه‌های مفیدی از تفکر درباره مسئله را فراهم می‌کند از جمله: تخصصی کردن مفهوم فرآیند کسب وکار به عنوان یک سرویس که توسعه‌ای توانمند شده از IaaS، PaaS و SaaS به وسیله محاسبات ابری می‌باشد. محاسبات ابری نقش مهمی را در این چارچوب دارد، با این حال فقط به عنوان یک عنصر در مجموعه‌ای از قابلیت‌ها دیده می‌شود که برای ارائه تحلیل آب و هوا به عنوان یک سرویس ضروری می‌باشد. این عناصر به این دلیل ضروری‌اند که در مجموع آن‌ها منجر به تولید ظرفیتی برای خود تجمیعی[۴] می‌شوند، که محققان این مقاله احساس می‌کنند که کلید حل بسیاری از چالش‌های داده‌های عظیم در این حوزه می‌باشد. سرویس تجزیه و تحلیل MERRA (MERRA/AS) نمونه ای از یک ابر CAaaS می‌باشد که بر اساس این اصل ساخته شده است[۸].

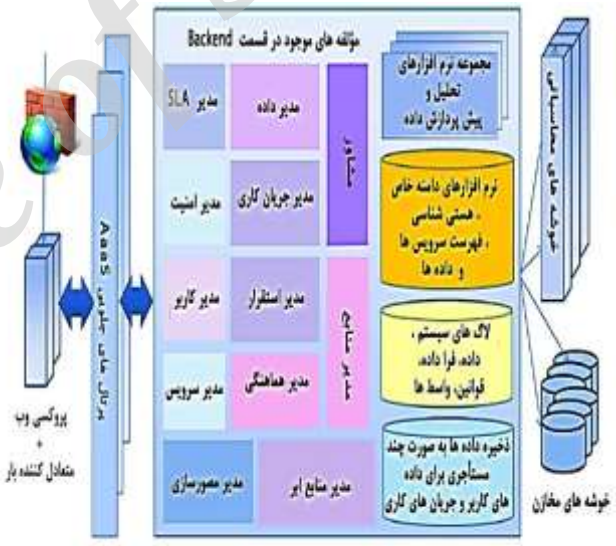
---

[۳] Schnase
[۴] Self-assembly

سازماندهی مقاله به این صورت است که در بخش دوم معماری مفهومی تحلیل مبتنی بر ابر به عنوان سرویس (CLAaaS) شرح داده شده است. در بخش سوم تکنولوژی هادوپ و کاهش نگاشت معرفی شده و در بخش چهارم روشی برای بهبود استفاد ه از هادوپ در CLAaaS ارائه گشته است و در نهایت، در بخش پنجم، نتیجه‌ی مختصری از موارد بیان شده ارائه می‌گردد.

۲- معماری مفهومی تحلیل مبتنی بر ابر به عنوان سرویس (CLAaaS)

در معماری مفهومی CLAaaS که در شکل (۱) آمده است، همانند هر سرویس ابر دیگر ترافیک اینترنت از دیواره آتش گذشته و به وسیلهٔ متعادل کننده بار و پروکسی وب هدایت و توزیع می‌شود. CLAaaS دارای دو قسمت واسط کاربر در جلو[۵] که شامل مجموعه‌ای از پرتال‌های وب سفارشی سازی شده برای دامنه‌های متفاوت و گروه‌های کاربری می‌باشد و قسمت موجود در پشت واسط کاربر[۶] که پرتال‌های محاسبات هسته‌ای در آن قرار دارند، می‌باشد. در قسمت راست نیز خوشه‌های محاسباتی و خوشه‌های مخازن که مخازن مقیاس‌پذیری می‌باشند که در هر زمانی لازم باشد، مورد استفاده قرار می‌گیرند، وجود دارند. مؤلفه‌های CLAaaS در سه بخش مدیریت سرویس، مدیریت جریان کاری و مدیریت داده دسته‌بندی می‌شوند.



شکل (۱): مدل مفهومی CLAaaS[۲]

۱) مؤلفه‌های مدیریت سرویس: مدیر توافق سطح سرویس[۷] (SLA)، مدیر امنیت، مدیر کاربر، و مدیر سرویس که وظایف مدیریتی سرویس را انجام می‌دهند. مدیر سرویس مسئول کارکرد و وضعیت کلی سیستم می‌باشد. نظارت بر درخواست‌ها، نگهداری اطلاعات نشست هماهنگی کارکردهای مؤلفه‌های مختلف به وسیله ارتباط کافی با آن‌ها نیز از وظایف این بخش می‌باشد. مدیر SLA مذاکرت SLA را برای سطوح مختلف سرویس‌ها و گروه‌های کاربری را پشتیبانی می‌کند و بر ایجاد SLAها نظارت دارد. مدیر کاربر به ثبت نام کاربران، نگهداری حساب‌های کاربران و تعریف سطح دسترسی گروه‌های کاربری مختلف کمک می‌نماید. امنیت سیستم توسط مدیر امنیت بر اساس سیاست‌های امنیتی از پیش تعریف شده و سیاست‌های خاص گروه‌های کاربری تأمین می‌شود.

۲) مؤلفه‌های مدیریت جریان کاری: مدیر داده، مدیر جریان کاری، مدیر مصورسازی[۸]، مدیر هماهنگ‌سازی، مشاور، مدیر منبع و مدیر منابع ابر که همگی با هم تشکیل سیستم مدیریت جریان کار را می‌دهند.

۳) مؤلفه‌های مدیریت داده: مجموعه نرم‌افزارهای تجزیه و تحلیل داده و سه منبع داده که در کل همگی وظایف پردازش و مدیریت داده را بر عهده دارند. مجموعه نرم‌افزاری می‌تواند شامل پردازش اختصاصی و سفارشی داده‌ها یا نرم‌افزار کاربردی پرس و جو، شامل داده‌های خارجی و یا سرویس‌های نرم-افزاری باشد. سه منبع داده اصلی در پلتفرم شامل موارد زیر می‌باشند:

- داده‌های مخصوص دامنه: شامل ابزار تحلیلی یکه دامنه خاص، اتصال به سرویس‌ها، قالب‌های جریان کاری، هستی‌شناسی، و منابع داده که فهرستی از تسهیلات استفاده مجدد و پرس و جو در آن وجود دارد.
- داده‌های مخصوص سیستم: شامل سیستم‌های کلی مرتبط با داده، فرا داده، لیستی از منابع ابر و کاربران جاری، سیاست‌ها و قوانین از پیش تعریف شده، لاگ‌ها عملیات سرویس و پرتال‌ها می‌باشد.
- داده‌های مخصوص کاربر: شامل داده‌های شخصی کاربران مثل کاربران ثبت نام شده یا SLAهای هر شخص، داده‌های بارگذاری شده یا لینک‌های داده تحلیلی، مدل‌های مصورسازی، مدل‌های تحلیلی و جریان‌های کاری می‌باشد.

۳- هادوپ[۹] و کاهش نگاشت[۱۰]

هادوپ یک ابتکار NoSql است، که قابلیت پشتیبانی از این حجم عظیم داده را دارا می‌باشد. ایجاد هادوپ به سال ۲۰۰۵ باز می‌گردد، که به عنوان ابتکاری از یاهو می‌باشد که از تکنولوژی کاهش نگاشت که قبلاً توسط گوگل به کار گرفته شده، استفاده می‌کند. از سال ۲۰۰۹ اکثر شرکت‌ها همچون گوگل، یاهو و فیس بوک برای جستجو، شاخص‌ها و سازماندهی از هادوپ استفاده می-کنند. هادوپ دارای فایل سیستمی توزیع شده به نام HDFS[۱۱] که در آن پکیج‌های نرم‌افزاری و ابزار گوناگونی وجود دارند که که داده‌های گوناگون از اکثر منابع از لاگ‌ها گرفته تا پایگاه داده‌های رابطه‌ای در آن ذخیره می-شوند[۳]. کاهش نگاشت روشی است که از محاسبات توزیع شده و پردازش موازی داده‌های عظیم را به وسیله کامپیوترهای بسیار قوی پشتیبانی می‌کنند. کاهش نگاشت محاسبات توزیع شده مجموعه بزرگی از داده‌ها را با تعداد زیادی از کامپیوترها (گره‌ها) انجام می‌دهد. در عملیات نگاشت گره اصلی[۱۲] (ریشه) ورودی را گرفته و به زیر مسئله‌های کوچکتر تقسیم کرده و هر زیر گره نیز تا جای ممکن این کار را تکرار می‌کند و در نهایت جواب‌ها به کاهش دهنده پاس داده شده و در این قسمت جواب‌ها با هم ترکیب شده و خروجی

[۵] Frontend
[۶] Backend
[۷] Service Level Agreement

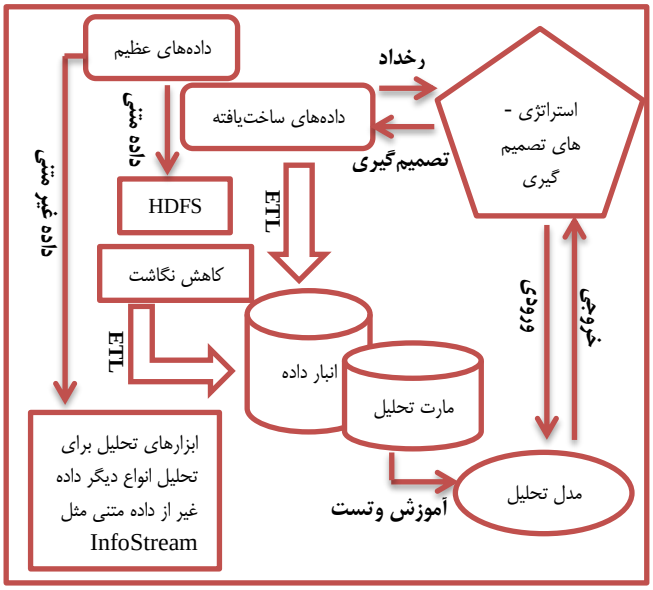
[۸] Visualization
[۹] hadoop
[۱۰] MapReduce
[۱۱] Hadoop Distributed File System
[۱۲] head

به دست می‌آید. یکی از مشکلات موجود در مورد هادوپ این است، تکنولوژی کاهش نگاشت موجود در آن فقط در مورد داده‌های متنی کاربرد دارد و بنابراین در استفاده از آن برای برخی از کاربردهای علوم که دارای داده‌های پیچیده چند بعدی و دودویی هستند، دارای محدودیت می‌باشیم [۸].

٤- روش پیشنهادی برای بهبود معماری CLAaaS

در مقاله [۶] از هادوپ به عنوان موتوری برای ETL استفاده شده است. ایده این روش تجمیع و یا ساخت متغیرهای عملکردی از داده‌های عظیم با استفاده از هادوپ می‌باشد در حالیکه داده‌های سنتی که نیاز به متعادل شدن و ترکیب شدن دارند، یک مسیر ETL معمولی را می‌پیمایند. یک مرتبه که داده‌های بزرگ به اندازه‌های قابل مدیریت کاهش می‌یابند، می‌توان با آن‌ها به عنوان داده‌های ساخت یافته معمولی که می‌توانند ذخیره شوند و در پایگاه داده‌ها پردازش شوند برخورد کرد. در این سناریو هادوپ شبیه به یک تکیه‌گاه و یک مؤلفهٔ ETL کارا عمل می‌کند.

با توجه به اینکه در معماری CLAaaS از انواع نرم افزارها همچون هادوپ و اینفواستریم[۱۳] برای تحلیل داده‌های عظیم استفاده می‌شود و همچنین به دلیل اینکه در برخی از زمینه‌ها ممکن است داده‌های هنوز به میزانی نرسیده باشند که بتوان به آن‌ها داده‌های عظیم گفت، مدل موجود در شکل (۲) برای ساخت‌یافته‌تر کردن قسمت تحلیلی معماری CLAaaS پیشنهاد شده است. در این پیشنهاد مدل معرفی شده توسط نعمان شیخ [۶] با اضافه کردن شرط-های داده‌های متنی و داده‌های غیر متنی به قسمت تحلیلی معماری CLAaaS اضافه می‌شود که با توجه به اینکه هادوپ یک ابزار قدرتمند NoSql می‌باشد در تمامی زمینه‌های داده‌های متنی به کار گرفته می‌شود و همچنین ظرفیت‌های موجود در مورد داده‌های معمولی که نیاز به اینگونه پردازش‌ها ندارند به کارگرفته نمی‌شوند واین موضوع موجب تسریع و بهبود روند در پردازش موازی و توزیع شده داده‌های عظیم می‌شود.



شکل (۲): روش ارائه شده برای بهبود معماری CLAaaS

[۱۳] InfoStream

٥- نتیجه‌گیری

تجزیه و تحلیل برای ایجاد یک دید بر روی داده‌های عظیم برای تصمیم-گیری مؤثر تعیین کننده می‌باشد. در این مقاله به معرفی معماری CLAaaS پرداخته شد زیرا این معماری از جریان‌های کاری تحلیلی برای ساختن پردازش‌های داده گوناگون و مرحله تحلیلی مورد نیاز برای بدست آوردن ارزش از داده‌های استخراج شده می‌پرداخت. سپس ساختار قسمت تحلیل این معماری را با مجزا کردن داده‌های متنی عظیم، داده‌های غیر متنی عظیم و داده‌های معمولی که نیاز به استفاده از ظرفیت‌های ویژه ایجاد شده ندارند، بهبود داده شد.

٦- منابع

[1] Haluk Demirkan , Dursun Delen*" Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud"* , Decision Support Systems, Volume 55, Issue 1, April 2013, Pages 412-421

[2] Zulkernine. F, Martin. P, Ying Zou, Bauer. M "*Towards Cloud-based Analytics-as-a-Service (CLAaaS) for Big Data Analytics in the Cloud"*, 2013 IEEE International Congress on Big Data

[3] Kim, H., Cho, I., and Yeom, H., 2008.*"A Task Pipelining Framework for e-Science Workflow Management Systems*"In IEEE International Symposium on Cluster Computing and the Grid, (CCGRID'08) pp. 657-662, IEEE.

[4] myExperiment workflow hosting site: available March 5, 2013 at: http://www.myexperiment.org/home.

[5] Dylan Maltby "*Big Data Analytics*", ASIST 2011, October 9-13, 2011, New Orleans, LA, USA.

[6] Nauman Sheikh "*Big Data, Hadoop, and Cloud Computing "*, Implementing Analytics, 2013, Pages 185-197

[7] H. Demirkan, R.J. Kauffman, J.A. Vayghan, H.-G. Fill, D. Karagiannis, P.P. Maglio, *"Service-oriented technology and management: perspectives on research and practice for the coming decade",* The Electronic Commerce Research and Applications Journal 7 (4) (2008) 356–376.

[8] John L. Schnase, Daniel Q. Duffy , Glenn S.Tamkin , Denis Nadeau , John H. Thompson, Cristina M. Grieg, Mark A. McInerney, William P. Webster *"MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service"* Computers, Environment and Urban Systems, In Press, Corrected Proof, Available online 31 January 2014