SBIR Phase I: A General-purpose Data Processing Pipeline for Knowledge Work Automation

Eric Griffis
EM Data Technology
eric@emdti.com

December 2, 2013

1 Identification and Significance of the Innovation

Enterprises of all sizes stand to benefit from access to innovative software technologies [8, 9, 20, 22]. Appropriate scale has, however, become a critical factor in the application of innovative software technologies toward operational efficiency optimization. For small and medium-sized businesses (SMBs), cost and awareness are substantial barriers to widespread adoption of advanced technologies [8, 14, 20]. This SBIR Phase I proposed project, dubbed *EM AutoFlow*, will advance the state of the art in knowledge work automation by leveraging modern programming language design principles to focus domain expertise onto shared service platforms in order to accelerate broad diffusion of innovation (DOI) in cutting edge, compute-intensive data processing technologies such as high-performance computing (HPC), machine learning (ML), and data mining (DM). By leveraging the well-studied benefits of elastic cloud and other shared service computing models [1, 17], the proposed technology will reduce costs substantially via appropriate economies of scale [5] and accelerate awareness of advanced solutions by delivery over ubiquitous channels like e-mail and the Web [14].

SMBs are intuitively the most likely to benefit from but least likely to pursue advanced automation and analytics solutions to business problems. According to the U.S. Small Business Administration (SBA), the 28 million SMBs that provided 99.7% of American jobs in 2010 generated $28.9 billion in revenue [19]. This important segment of the American economy includes various forms of high risk but minimally funded operations for which process optimization can have profound benefits to financial stability at every level, from individuals employed by SMBs to the economies in which SMBs operate.

For large enterprises, middleware services [3] were the de-facto solution to scalable software automation for decades, but standardization of software APIs, protocols, and frameworks are sufficient only for enterprises with the means to acquire and maintain substantial software engineering talent. Service oriented architecture [12] (SOA) alleviates some of the burden of middleware adoption by masking low-level technical details behind standard message-passing facilities, but SOA does not eliminate the need for in-house software engineering talent. Business and scientific workflow technologies [2] do eliminate completely the engineering complexities involved in data-intensive processing tasks. Although this approach has been successful at decoupling specific classes of users and applications from the underlying technical expertise, implementation divergence and infrastructural complexity management remain prominent barriers, as evidenced by the emergence of web-based, general-purpose management systems [10].

Each of the above solutions targets end users of a distinct level of technological expertise without addressing directly the core usability issues relevant to basic business needs; indeed, substantial a priori knowledge of the relationship between high-level needs and any supporting tools is assumed. The EM AutoFlow design will facilitate direct transformation of high-level business needs into valuable data products by providing domain experts with a simple, general-purpose environment in which to project arbitrary special-purpose technologies with distinct infrastructural requirements into intricate, re-usable compositions.

1.1 Broader impact

The benefits of the proposed innovation are not limited to SMBs. Large businesses stand to benefit internally from access to a robust and mature extreme-scale software automation and integration platform, especially when combined with quality consulting services expected to emerge from the concentrations of highly specialized expertise encouraged by proliferation of the EM AutoFlow platform. Through intelligent application of the proposed technology, overhead costs associated with regulatory pressures can largely be distributed among cooperative businesses by the same mechanisms that reduce other costs for SMBs. Large academic and government efforts likewise will be more capable of dealing with arbitrarily complex requirements on resource tracking, reproducible workflow behavior, provenance tracking, non-disclosure enforcement, and other mundane but important operations not driving primary outputs directly.

Society at large stands to benefit measurably as well. The cost-distributive nature of shared service platforms can extend DOI to individual citizens in the form of subscription-based services geared towards personal software automation and analytics. By nature of the proposed technology, innovative services that leverage the EM AutoFlow platform will be well-positioned to participate in inter-enterprise collaboration of unprecedented proportion. Municipal efforts can also apply the proposed platform to advance internal operational efficiency to affect improvements in tax revenue expenditures, new avenues for civil engagement and transparency, inter-entity collaborations, and other activities presently considered unfeasible due to communications overhead complexity and inefficient distribution of expertise. The proposed innovation may also carry over to markets in developing countries, where extreme scarcity of talent and other resources amplify many of the needs and obstacles outlined above.

2 Background and Phase I Technical Objectives

The driving technical objective of this project is to derive a high-level programming environment that strikes an appropriate balance between expressiveness and ease of use with respect to data-intensive software process automation. Commercial feasibility is, however, the major factor in identifying a desirable balance and is of primary import. Intuitively, the complexity of automating a software system grows in direct proportion to the sophistication of the components to be automated and so there exists a point at which the cost of automation outweighs any potential benefits. The proposed technical and commercial objectives are therefore coupled tightly. This tension is perhaps under-emphasized by the existing body of related work.

2.1 Case study

To focus the research effort, a case study on small-scale Accounting process automation will be developed. The resulting design, named *AutoAP*, will capture automatic invoice processing, transaction recording, and financial report generation. A small group of volunteer SMBs will be recruited in order to obtain data representative of realistic workloads. Accounting is an ideal subject for case study for several important reasons. Each of the tasks targeted for automation can benefit from some amount of intelligent coordination or heavy computation for which effective use is currently a clear financial burden. Furthermore, the typical

SMB accounting process is manual and tedious, so demand for affordable automation will likely grow quickly once available. Finally, the EM AutoFlow R&D team possesses Accounting expertise. If the case study is successful, then the following lists of technical and commercial objectives will serve as templates for iterative exploration of subsequent niches, markets, and industries.

2.2 Technical feasibility

The key objectives specific to technical feasibility are enumerated here, followed by detailed explanations. The technical objectives emphasize efficient capture and representation of relevant inputs and subsequent processing toward useful outputs.

1. Identify computing tasks suitable for automation.

   The first step toward meaningful production will be to uncover basic technological needs of the target market. What business needs are currently manual, insufficiently automatic, or cost prohibitive for SMBs in the target market? Can these tasks benefit from advanced technological solutions? If so, what major factors preclude existing solutions? The R&D team will rely on target market domain expertise to produce the initial list of potential tasks for advanced automation and analytics. With these tasks in mind, relevant technological products can be derived, such as hardware or software data sources, data processing applications, and supporting hardware infrastructure.

2. Determine primitive use cases and representative composite analogues.

   After the initial list of potential resources has been drafted, the next step will be to synthesize expected use cases and identify any complex constructions with simpler underlying primitives. How are the tasks underlying use cases similar or different? Are any patterns observable within or between constructions? The primitive constructs are ideal candidates for early abstraction, as they are likely to capture basic deployment requirements and be re-used most frequently.

3. Design high-level programming facilities and low-level deployment mechanisms.

   Primitive abstractions will have to be wired into a top-level environment via domain-specific languages (DSLs) along with hints for optimal deployment such as input and output data formats, factors affecting parallelism, and regulatory requirements. What sort of language constructs are required to express the sum of requirements in the simplest manner possible? Which families of tasks are inherently the most difficult to express? The Basic DSL mechanisms will be designed according to the requirements of the case study. Overall technical feasibility will largely be influenced by successful crafting of appropriate primitives and DSL-construction facilities. Low level tasks like job scheduling, message passing, and data storage and marshaling will have to be handled by existing technologies. Which technologies are available to satisfy the identified requirements? If multiple options exist, what are the trade-offs? Can these trade-offs be expressed, inferred, or differentiated by structured or unstructured automatic learning processes?

4. Analyze computational requirements.

   The implementation details for representative use cases will be considered in a comprehensive requirements analysis that involves identification and measurement of all details related to computational costs. How much computational power is required for each task? Which tasks are most expensive in this regard? Can more expensive tasks be substituted or approximated by less expensive ones? If so, can situations in which substitution or approximation is appropriate be inferred systematically or specified directly? Are there any opportunities for automatic process optimization? The various metrics produced by previous objectives will be combined with raw performance benchmarks, reproducible results of community- and vendor-conducted research, and asymptotic algorithm analysis to predict run-time costs in terms that can readily be converted into rates of resource consumption such as computational cycle requirements, suitability for various approaches to parallelization, sustained I/O capabilities, temporary and persistent storage requirements, and electrical power draw.

5. Verify sub-linear scaling of domain expertise with respect to computational requirements.

   Conclusive technical feasibility is determined directly by whether the combined result of all previous technical objectives admit sufficient scaling of human expertise. Given the sum results of previous objectives, which particular solutions make sense to implement, and in what order? Does the high-level environment design allow for rapid and dynamic adaptation to evolving end-user needs? According to the experiences of the R&D team, rapid invention and adoption of advanced technologies can enable computational resource scaling proportional to the logarithm of applied human expertise. In other words, as demand for domain expertise increases linearly, manageable computational power can reasonably be expected to increase exponentially for some time before saturating. Validation of this hypothesis and accurate prediction of the saturation point will decide the technical feasibility of the designs produced for a given target market.

2.3 Commercial feasibility

The key objectives specific to commercial feasibility are enumerated here, followed by detailed explanations. The commercial objectives emphasize cost-per-value comparisons of both the design of the high-level operating environment as well as the process of translating business needs into data products.

1. Identify relevant market opportunities.

   The first step toward effective commercialization will be to observe opportunities for advanced automation and analytics in new markets. What problems currently affect SMBs in the target market? Do acceptable solutions already exist for these problems? The primary goal of this objective is simply to inform the pursuit of relevant technical objectives and will not extend into comprehensive market research. Although Accounting, for instance, has already been selected as the target market for case study, various details require brief consideration in order to justify commercial feasibility, such as confirmation of actual needs by literature review, presence or absence of related technological solutions, or current job postings. Conditions and willingness of case study volunteers to commit will also be considered.

2. Characterize representative business needs.

   Specific high-level business needs will be defined in order to motivate both sides of the feasibility equation. What sort of business operations are typical to the target market? Which aspects can benefit from advanced automation and analytics? To what degree and in what order are solutions appropriate? Identifying non-technical needs will impart form to the use cases that drive technical feasibility objectives by putting into context the relevance of specific existing technologies. Commercial feasibility and the degree of innovation clearly depends on the ability for the proposed technologies to solve real problems. Moreover, SMBs are more likely to experience the need for advanced solutions in a high-level, top-down manner motivated by economic pressures as opposed to the detail-driven, bottom-up fashion assumed by existing solutions. Can EM AutoFlow be applied to the transformation of high-level business needs into valuable data products and services?

3. Design reasonably effective interfaces for non-expert users.

   With high-level needs and results in mind, reasonably effective interaction patterns and supporting interfaces for non-experts will be designed and tested. How much information is enough to deliver the desired data products and services? What are possible strategies for capturing existing information, generating missing information, and delivering end results? What degree of technical proficiency can reasonably be expected of the typical end user? What sort of atypical usage patterns can be anticipated, and can the needs of atypical users be addressed efficiently? Because EM AutoFlow must be designed to reduce complexity, special care will be taken to deliver an intuitive experience to all parties involved. As the levels of underlying complexity increase, so will the demand for expertise at all levels of design and implementation. Can EM AutoFlow be designed to constrain user interface complexity in addition to architectural and implementation complexity?

4. Analyze overall cost and compare against market rates for comparable work.

   Though conclusive commercial feasibility will not be a primary commercial objective of Phase I, the combined results of these commercial objectives can give preliminary insights into justification of the costs associated with the Phase I technical objectives. What are the secondary costs associated with construction and maintenance of an EM AutoFlow instance? To what degree are basic requirements like electrical power, redundancy, secure backups, internal provenance tracking, and long-term data archival of practical concern? Can EM AutoFlow offer competitive alternatives to existing business solutions? Given the form of commercial opportunities, adequate end-user experience elements, and technical details, a simple cost-per-need prediction will suffice to justify comprehensive market research in a later phase.

3 Phase I Research Plan

The primary innovation of EM AutoFlow is the introduction of general-purpose programming environments suitable for modern knowledge work automation tasks. Existing approaches to automation tend to impose substantial constraints on operational effectiveness, including an assumptions of access to diverse expert talent or of extremely narrow user and application bases. The EM AutoFlow programming environment will be designed to deliver intuitive interfaces for overcoming these limitations at every level of detail. Specifically, a successful knowledge work automation environment must allow for the natural expression of high-level knowledge work tasks, low-level supporting tasks, arbitrary compositions of advanced domain-specific technologies, resource optimization, intelligent deployment strategies, and end-user interactions as outlined in figure 1.

The research effort will be carried out in a three-part sequence. The first part consists of a two- to three-month study into the relevant use cases and existing technologies resulting in designs that capture the underlying conceptual details. For the second part, the R&D team will spend two months constructing and iteratively testing and tweaking models of the various concepts in order to collect the measurements necessary for determining feasibility. In the final part, the team will require three to four weeks to perform a detailed analysis of the results and assemble their findings into a comprehensive technical report.

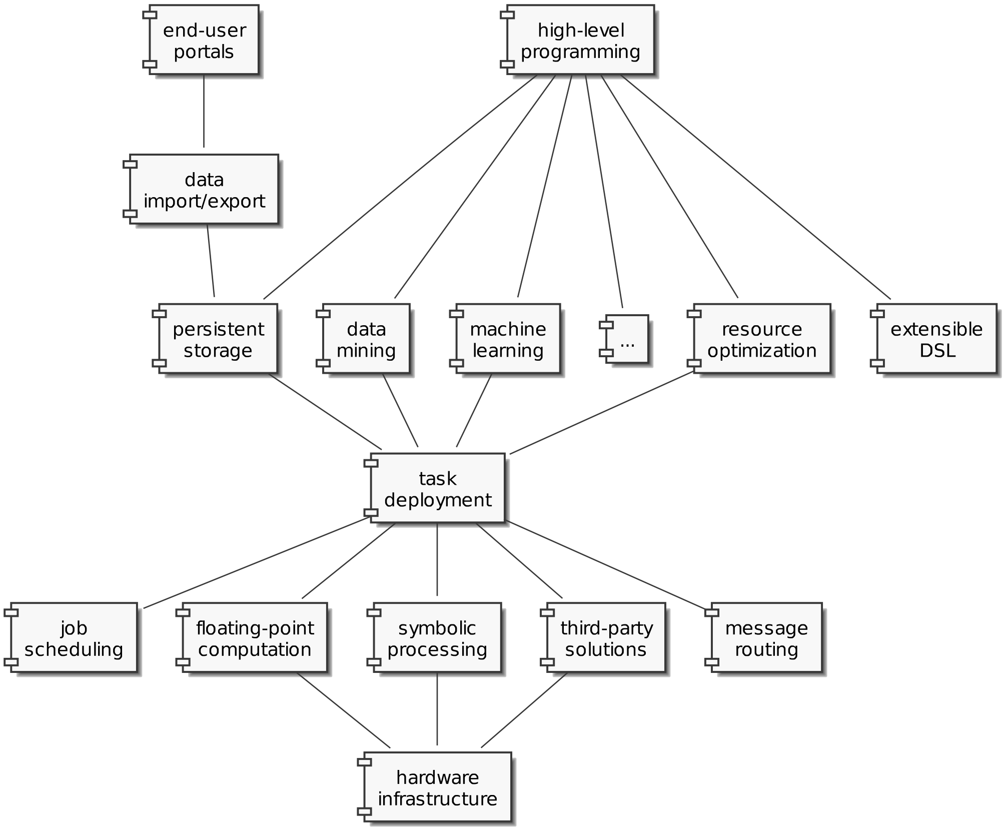

Figure 1: EM AutoFlow component architecture.

3.1 Concept designs

Concept designs will be required for elements of end-user interfaces, the high-level programming environment, lower-level computational models, and underlying hardware infrastructures. The R&D team will conduct research into concept designs in the following sequence.

1. Create a list of candidate high-level tasks.

   Discover and review academic and trade publications related to Accounting, technologies supporting operations at Accounting firms of various sizes, and the enterprises that employ the services of those firms. Note apparently useful combinations of tasks and advanced technologies.

2. Classify existing related solutions.

   Rank each of the findings according to accessibility by SMBs with respect to purchase price, combined expected setup and maintenance costs, operational restrictions, and publication target demographics. Purchase price is the average published price. Setup costs include minimum hardware and software requirements. Maintenance includes consumables like power, network bandwidth, and storage media. Operational restrictions include access to resources like supercomputers, data centers, and in-house technical expertise.

3. Characterize nature of and potential uses for advanced technologies.

Match high-level tasks to opportunities for broad application of technologies like stream processing [11], distributed storage [21] and processing [7, 15], logic programming [4, 6], and data analytics [13, 18]. Categorize tasks according to technological requirements.

4. Formulate concepts for application of advanced technologies to specific tasks.

   Parameterize categories to various levels of approximation of the tasks that inhabit each category. Produce syntactic and semantic abstractions that generalize any similarities or differences observable among the parameter sets. Produce user interface requirements that impose end-user intuition on these abstractions such as familiar parameter names and compositions with analogies to common real-world phenomena.

5. Identify data sources and sinks.

   Begin ongoing collection and organization of input data from case study volunteers. Survey case study volunteers for desired output data. List internal and external data sources and sinks inherent in specific tasks, task categories, and technologies. Examples of external data sources include Web forms, digital cameras, document scanners, barcode scanners, RFID scanners, digital signature pads, biometric scanners, and credit card readers. Examples of external data sinks include Web portals, FTP dumps, outbound e-mail servers, printers, SMS messages, and social media outlets. Various architectural components and integrated third-party software systems comprise the internal data sources and sinks.

6. Identify supporting technologies.

   In Phase I, supporting technologies will consist entirely of standard open source operating systems, libraries, and programming languages; e-mail and Web 2.0 software stacks; and Intel-based commodity hardware.

7. Discover opportunities for systematic optimization.

   Observe apparent preferences of specific task categories for particular technologies and parameter configurations. Attempt to encode the observed preferences as logic programs.

8. Derive theoretical limitations.

   Determine the theoretical maximum workloads possible on available Phase I supporting commodity hardware by applying standard formulas to published hardware specifications. Compare to the published algorithmic complexities of potentially useful advanced technologies.

3.2 Modeling

Concepts will be implemented crudely to facilitate correctness and cost testing. Primary operational costs will be decided by input and output performance as well as resource consumption. The R&D team will test the results of the first part according to the following cycle.

1. Devise correctness and benchmark tests.

   Produce some desired case study output data manually. Implement test harnesses to verify the integrity of input and output data. Establish bounds on acceptable performance according to technological limitations and high-level workload demand. Instrument the test environment to produce time series utilization of CPU and RAM, and throughput of I/O channels. Estimate artificial resource demands imposed by the test environment and instrumentation.

2. For each primitive concept and frequent combination of primitives:

   - (Re-)implement the model.
   - Verify process and result correctness.
   - Conduct and record benchmarks.
   - Adjust the model.
   - Repeat.

   Produce a minimally functional implementation of the concept or concepts to be tested. Note any readily correctable deficiencies in algorithm selection and implementation. Reset the testing environment to a consistent initial state. Inject a uniformly random sample of data along with the model implementation into testing environment and perform a test execution. Continue resetting the testing environment and performing executions on various sample sizes to obtain statistically significant results. If the results clearly suggest improvements to the minimally functional implementation, adjust the implementation and re-iterate with new sample data. Repeat this cycle until results are conclusive or time allocated for testing the model expires.

3. Document interesting interactions.

   Observe pronounced synergistic behavior in desirable areas such as ease of use, conceptual intuition, and performance. Document confirmation or refutation of expected results.

3.3 Analysis & reporting

In the final weeks of the project, the results collected in the two previous parts will be analyzed and presented in a comprehensive technical report.

1. Present quantitative results.

   Plot performance versus time for each model and performance versus model for each high-level task. Estimate per-model primary costs. Estimate minimum per-task primary costs. Outline primary cost and benchmark trends. Incorporate setup and maintenance factors into minimum per-task cost estimates. Explain interesting results, including notable difficulties encountered during the design and modeling efforts.

2. Predict model and task scalability.

   Extrapolate scalability trends from the quantitative results for each model and task according to saturation point—roughly 2.5 standard deviations from expected capacity versus operational overhead. Derive a defensible threshold for classifying concepts as successful based on predicted scalability of associated models and tasks.

3. Confirm technical and commercial feasibility.

   Confirm technical feasibility by presenting systematic formulations of concept as formal grammars and operational semantic rule sets. Evaluate the apparent strengths and weaknesses of the final design with respect to simplicity, performance, and ease of use. Confirm commercial feasibility with respect to case study by comparing estimated total cost to implement and maintain the final design versus existing comparable solutions at optimal scales. Note failures to deliver promising or desired solutions. Document opportunities for future work.

4 Commercial Potential

4.1 The market opportunity

In the long term, various markets that employ knowledge work will be targeted—e.g., finance, law, medical records, engineering, Internet service providers, and high-tech manufacturing—but the initial anticipated target market is SMBs. Specifically, American SMB Accounting service needs will be targeted initially. For most industries, the SBA defines a "small business" in terms of the average number of employees and annual receipts. Finance is a key component of every business—every business needs to pay bills—so SMBs inevitably will seek technology to stay competitive. The potential customer will own or operate a small to medium-sized business with specific Accounting service needs according to standard business finance and accounting principles. The customer will likely have need for Accounting software, planning and time tracking tools, communications platforms, or mobile access to information.

The proposed innovation will address customer needs by reducing substantially the time, cost, and uncertainty associated with routine business operations. The EM AutoFlow platform will replace manual data entry with digital capture devices like digital scanners and biometric sensors, likely via mobile device. High-level programs written by an Accounting expert, possibly with assistance from a programmer advocate, will process the data and perform routine calculations. The platform will provide outsourced data processing and management services at previously unprofitable scales via a shared service approach with elastic demand response, providing confidence in daily operations via advanced reports, alerts, automatic pattern discovery, built-in disaster recovery mechanisms, and secure remote access to information with intuitive third-party access control.

The initial target market size is on the order of millions or tens of millions of distinct businesses domestically. According to the SBA, the majority of sales and jobs in America currently are generated by 28 million SMBs [19]. In the long term, the target market size is likely much larger. The McKinney Global Institute has estimated global economic potential of knowledge work automation technologies in excess of 5.2 - 6.7 trillion dollars by the year 2025 [16] and point to the pervasive enabling nature of information technology as an indication that this estimate is conservative.

The most prominent barrier to entry will be the overhead associated with achieving appropriate economies of scale. Though some classes of advanced technology—e.g., super computer simulations—clearly operate most efficiently at massive scale, research suggests this is not the case for some common but important classes [5]. These common cases will be targeted first.

Competing products and services will, however, pose an immediate barrier to entry. On one hand, the business process outsourcing (BPO) space is already inhabited by cost-heavy solutions from big names like IBM and Accenture. The EM AutoFlow brand must highlight differences between the nature and scope of its services and those offered by, e.g., IBM Global Process Services. On the other hand, the brand must also set itself apart from established desktop and cloud-based offerings like Quickbooks, Xero, and Freshbooks, which target individuals and small businesses but lack the flexibility and sophisticated automation provided by EM AutoFlow.

4.2 The Company/Team

EM Data Technology is a family-owned and operated business that grew from fifteen years of entrepreneurial exploits in diverse areas like Accounting, business consulting, financial services, data mining, Computer Science research, high-tech consulting, and Internet service provision. The company consists of two employees, supported by outside design, engineering, and marketing talent contracted as needed.

The principal is Eric Griffis, Chief Scientist. Eric sets the company's direction of technological developments and manages R&D and engineering efforts. He brings to the project expertise in large scale software automation and tool construction. Eric received the B.S. in Mathematics of Computation from University of California, Los Angeles (UCLA) and has been recognized by the California Senate for outstanding achievements in STEM. He has worked at top domestic and international research institutions and has published work in programming languages and scientific workflow automation at a top conference. Eric expects to receive the M.S. in Computer Science from UCLA in June 2014. Before entering higher education, Eric spent a decade in ISP systems automation, data mining and curation, deep space instrument and spacecraft navigation operations automation at NASA JPL, analytic tool construction, software engineering, system administration, technical support automation, and infrastructure architecture and automation R&D for Internet service providers through various stages of growth.

The co-principal is Mei Griffis, Operations Director. Mei oversees the company's business and financial strategies and directs daily operations. She brings to the project expertise in Accounting and enterprise integration. Mei received the B.S. in Finance from the University of Southern California, the M.A. in Education from California State University, Los Angeles, and is a Certified Public Accountant. She maintains affiliations with the AICPA and CALCPA organizations. Relevant work experience includes business management and consulting, large scale business systems analysis and integration, Accounting including accounts payable and auditing for firms of various sizes, technical training, project management, database administration, ERP, and regulatory compliance.

As a result of this innovation, the company anticipates substantial impact on job creation at each of the major levels of the component architecture presented in figure 1. The core platform will maintain the core R&D team to follow subsequent advances in knowledge work automation and a software engineering team to translate the results into robust platform features. To maintain relevant subscription service offerings, each target market will need a domain expert and possibly a programmer advocate to apply new features and keep up with evolving business needs. Marketing and sales expertise will be required to extend subscription-based services into new target markets, as well as to pursue large enterprise partnerships and affiliate relationships.

EM Data Technology was created when its founders, having identified various data products that could improve society measurably, recognized that a growing capacity for data collection would promote demand for data products that outstripped processing capacity. The company's mission, then, is to identify and pursue opportunities to promote widespread diffusion of innovation in automated data processing. In this regard, the proposed research meshes well with the company's existing objectives and provides a vehicle for long-term sustainable innovation that sits squarely within the company's mission.

4.3 Product or technology and competition

EM AutoFlow will face competition primarily from existing product vendors and service providers. Though product and services similar to EM AutoFlow do exist, outright competition is mitigated by two facts. First, the innovation will provide vastly superior service at costs comparable to desktop and cloud-based alternatives. Second, the innovation will create truly new opportunities for SMBs to enjoy many of the benefits of previously cost-prohibitive solutions. EM AutoFlow will continuously leverage advanced knowledge work automation to strike a new balance between high- and low-end alternatives that can deliver more features while lowering cost and turn-around time. The innovation will enjoy an initial period with no direct competition in which to dominate the new space. During this initial period, corporate partnerships and affiliate programs will be pursued aggressively to mitigate any potential for direct competition. To state the value proposition explicitly, *services enabled by the EM AutoFlow innovation will leverage advanced knowledge work automation technologies to automate routine business operations, which will save time and money and* *eliminate uncertainty and sources of human error.*

EM Data Technology identifies proprietary intellectual property (IP) as an obstacle to its mission of maximum diffusion of innovation. The company believes that truly disruptive software is enabled by discovery of radically simple concepts and implementations, and that rapid software duplication is inevitable. The company also believes that integrity is critical to long-term viability and that the real market value of disruptive automation technology emerges from the research, development, support, and community-building efforts that produce it. Thus, the most effective way to protect IP is to publish results in peer-reviewed journals and to release research prototypes as open-source projects with licenses that protect contributors from liability and exploitation.

Critical milestones that must be met to bring EM AutoFlow to market are as follows.

1. Produce a feasible platform design.
2. Build a prototype platform.
3. Acquire talent to support a pilot service.
4. Initiate a pilot service in the target market.
5. Update the platform design according to pilot service findings.
6. Acquire talent to support full scale roll-out.
7. Implement a robust platform.
8. Complete initial target market roll-out.

The pilot service will require legal services and additional engineering capacity. Implementation of a robust platform will require further engineering capacity. Initial roll-out will conclude by ramping up support, sales, and marketing activities.

4.4 Financing and revenue model

Subscription Plan and Revenue Forecast

The standard sales forecast model is presented in figure 2. It is suggested that if 5000 new subscriptions occur during the first month of business operation, the company will generate $149,750 which equals $1,797,000 annually.

The revenue assumption is based on a 3 tier subscription-based pricing model. The three tiers are Starter $29.95 per month, Professional $59.95/month, and Corporate $99.95/month. An Enterprise tier would be also provided with a custom quote. Levels to be provided for online subscription-based services will be as detailed in figure 3.

Revenue	Fee Basis/month	Conservative	Standard	Optimistic
		1000 subscriptions	5000 subscriptions	10000 subscriptions
Monthly	$29.95	$29,950.00	$149,750.00	$299,500.00
Yearly		$359,400.00	$1,797,000.00	$3,594,000.00

Figure 2: AutoAP basic revenue forecast.

	Starter	Professional	Corporate
	$29.95	$59.95	$99.95
Bill Processing	✓	✓	✓
Web Portal	✓	✓	✓
Bill Tracker	✓	✓	✓
Auto Payments	✓	✓	✓
Reports		✓	✓
Document Scanner		✓	✓
Bookkeeping + Reconciliation		✓	✓
CPA Service			✓
2000+ Reports Available			✓
Advanced Web Portal			✓
High-Volume Process			custom quote

Figure 3: AutoAP predicted levels of service.

In our conservative sales forecast model, the break-even unit lies around 1000 subscribers. If there are 1000 subscribers to the AutoAP “Starter” plan in the first month, the company will generate $29,950 cash revenue equal to $359,400 for the first year. Assuming the company begins at this level of operation, the operating expense will be approximately $300,000 per year according to the Phase I budget analysis. This basic figure would give a net profit of $59,400 for the first year. Please be aware that subscription rates of less than 1000 subscribers would likely lead to a loss for the company. A more accurate figure will be available after the Phase I feasibility study concludes.

Revenue categories are generated by subscription-based services for end users, consulting services for enterprise partners, and affiliate programs such as accounting and business processing outsourcing providers that deploy AutoAP technology in their own Accounting firms.

Financing Plan

Before considering outside funding, the company will ask several questions during SBIR Phase II: is the innovation prototype mature enough to enter the market? What amount of subscription revenue would allow the company to self-fund further operations? If outside funding is actually needed, how much does the company need and what would be the terms?

Due to the nature of software systems, a viable product might be available for use by the end of Phase II. If this occurs, working capital could be provided by a small business loan from the Small Business Administration (SBA) specifically for minority-owned businesses. According to the SBA website, 7(a) loans have a maximum loan amount of $5 million. SBA does not set a minimum loan amount. The average 7(a) loan amount in fiscal year 2012 was $337,730. The interest rate of an SBA loan is relatively reasonable as compared to Small Business Investment Companies (SBIC), which will be mentioned below. The company maintains a good relationship with an experienced business banker who manages both SBA loans and business financing at JP Morgan Chase Bank N.A.

The SBIC program, licensed and regulated by the SBA and listed member of the National Association of Small Business Investment Companies (NASBIC) and National Association of Investment Companies (NAIC). is another option geared toward financing for minority-owned business. Mei was an asset to her previous employer, 24 billion asset under management (AUM) alternative investment firm Kayne Anderson Capital Advisors. She contributed to many technological projects. High interest rates might be a barrier to obtaining these loans in absence of a complete understanding of the growth trends of the business, which might not be certain for several years.

Venture Capital (VC) funding would likely be the last option. With VC funding, business strategy may be influenced by VC firms. There are no significant benefits to the company for obtaining VC funding and, more importantly, project success is not guaranteed with high capital funding. Therefore, VC funding is less desirable. Without VC funding, ownership will be kept 100% within the business and the company will have more freedom to operate.

5 Consultants and Subawards/Subcontracts

5.1 Consultants

The project will potentially hire a paid Machine Learning Engineer to assist in identifying opportunities to apply machine learning algorithms to all aspects of the project, and to assist in designing concepts and implementing models that involve machine learning. Potential candidates must possess a degree in Computer Science, Applied Math, Information Theory, or a related field.

5.2 Subawards

NONE

6 Equivalent or Overlapping Proposals or Awards to/from Other Federal Agencies

NONE

7 Lineage of the Innovation

The proposed work has connections to the following NSF award:

CSE Directorate for Computer & Information Science & Engineering
Division of Advanced CyberInfrastructure
Award 0123937

References

[1] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, and Ion Stoica. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, 2010. URL http://dl.acm.org/citation.cfm?id=1721672.

[2] Adam Barker and Jano Van Hemert. Scientific workflow: a survey and research directions. In *Parallel Processing and Applied Mathematics*, page 746–753. Springer, 2008. URL http://link.springer.com/chapter/10.1007/978-3-540-68111-3_78.

[3] Philip A. Bernstein. Middleware: a model for distributed system services. *Communications of the ACM*, 39(2):86–98, 1996. URL http://dl.acm.org/citation.cfm?id=230809.

[4] Stefano Ceri, Georg Gottlob, and Letizia Tanca. What you always wanted to know about datalog (and never dared to ask). *Knowledge and Data Engineering, IEEE Transactions on*, 1(1):146–166, 1989. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=43410.

[5] Kenneth Church, Albert Greenberg, and James Hamilton. On delivering embarrassingly distributed cloud services. *Hotnets VII*, 34, 2008. URL http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.150.606.

[6] Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient SMT solver. In *Tools and Algorithms for the Construction and Analysis of Systems*, page 337–340. Springer, 2008. URL http://link.springer.com/chapter/10.1007/978-3-540-78800-3_24.

[7] Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008. URL http://dl.acm.org/citation.cfm?id=1327492.

[8] Robert G. Fichman. Real options and IT platform adoption: Implications for theory and practice. *Information Systems Research*, 15(2):132–154, June 2004. ISSN 1047-7047, 1526-5536. doi: 10.1287/isre.1040.0021. URL http://pubsonline.informs.org/doi/abs/10.1287/isre.1040.0021.

[9] Robert G Fichman and Chris F Kemerer. Adoption of software engineering process innovations: The case of object orientation. *Sloan Management Review*, 1993. URL http://www.pitt.edu/~ckemerer/CK%20research%20papers/AdoptionSwEngineeringProcessInnovation_FichmanKemerer93.pdf.

[10] Sandra Gesing, Peter Kacsuk, Malcolm Atkinson, Iraklis Klampanos, Michelle Galea, Michael R. Berthold, Roberto Barbera, Diego Scardaci, Gabor Terstyanszky, and Tamas Kiss. The demand for consistent web-based workflow editors. In *WORKS '13 Proceedings of the 8th Workshop on Workflows in Support of Large-Scale Science*, pages 112–123. ACM Press, 2013. ISBN 9781450325028. doi: 10.1145/2534248.2534260. URL http://dl.acm.org/citation.cfm?doid=2534248.2534260.

[11] Jim Gray and Andreas Reuter. *Transaction processing*. Kaufmann, 1993. URL https://silk-weaver.googlecode.com/hg-history/21d238b0a409be8ab2d0ed428e3358019ac76d29/doc/spec/silk-spec.pdf.

[12] Hao He. What is service-oriented architecture. *Publicação eletrônica*, 30, 2003. URL http://www.nmis.isti.cnr.it/casarosa/SIA/readings/SOA_Introduction.pdf.

[13] IT Jolliffe. Principal component analysis. *Springer Series in Statistics, Berlin: Springer, 1986*, 1, 1986. URL http://hbanaszak.mjr.uw.edu.pl/MarketingoweZastosowania/PCA/Jolliffe_2002_PrincipalComponentAnalysis.pdf.

[14] J. W. Long. Lorenz: Using the web to make HPC easier. Technical report, Lawrence Livermore National Laboratory (LLNL), Livermore, CA, 2013. URL https://e-reports-ext.llnl.gov/pdf/760475.pdf.

[15] Grzegorz Malewicz, Matthew H. Austern, Aart JC Bik, James C. Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski. Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, page 135–146, 2010. URL http://dl.acm.org/citation.cfm?id=1807184.

[16] James Manyika, Michael Chui, Jacques Bughin, Richard Dobbs, Peter Bisson, and Alex Marrs. Disruptive technologies: Advances that will transform life, business, and the global economy. *McKinsey Global Institute, May*, 2013. URL http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Disruptive%20technologies/MGI_Disruptive_technologies_Full_report_May2013.ashx.

[17] Sean Marston, Zhi Li, Subhajyoti Bandyopadhyay, Juheng Zhang, and Anand Ghalsasi. Cloud computing — the business perspective. *Decision Support Systems*, 51(1):176–189, April 2011. ISSN 01679236. doi: 10.1016/j.dss.2010.12.006. URL http://linkinghub.elsevier.com/retrieve/pii/S0167923610002393.

[18] Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and K. R. Mullers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, page 41–48, 1999. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=788121.

[19] SBA Office of Advocacy. Frequently asked questions about small business, September 2012. URL http://www.sba.gov/sites/default/files/FAQ_Sept_2012.pdf.

[20] Mark S Shephard, Cameron Smith, and John E Kolb. Bringing HPC to engineering innovation. *Computing in Science & Engineering*, page 16–25, 2013. URL http://www.scorec.rpi.edu/REPORTS/2012-1.pdf.

[21] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, page 1–10, 2010. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5496972.

[22] Constantinos J Stefanou. Adoption of Free/Open source ERP software by SMEs. In *Information Systems for Small and Medium-sized Enterprises*, page 157–166. Springer, 2014. URL http://link.springer.com/chapter/10.1007/978-3-642-38244-4_8.